ARL-TR-8277 ● Jan 2018

US Army Research Laboratory

Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness (UMMPIREE): Phase I Report

by Clayton Burford, Lauren Reinerman, Grace Teo, Gerald Matthews, Joseph McDonnell, Kara Orvis, Mark Riecken, Peter Hancock, and Christopher Metevier

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8277 ● Jan 2018

US Army Research Laboratory

Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness (UMMPIREE): Phase I Report

by Clayton Burford and Christopher Metevier
Human Research and Engineering Directorate, ARL

Lauren Reinerman, Grace Teo, Gerald Matthews, and Peter Hancock
Institute for Simulation and Training, University of Central Florida, Orlando, FL

Joseph McDonnell
Dynamic Animation Systems Inc., Fairfax, VA

Kara Orvis
Aptima Inc., Orlando, FL

Mark Riecken
Trideum Corporation, Huntsville, AL

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* January 2018	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)** 25 April 2016–25 April 2017

4. TITLE AND SUBTITLE Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness (UMMPIREE): Phase I Report	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Clayton Burford, Lauren Reinerman, Grace Teo, Gerry Matthews, Joseph McDonnell, Kara Orvis, Mark Riecken, Peter Hancock, and Christopher Metevier	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Research Laboratory 12423 Research Parkway Orlando, FL 32826	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-8277
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) US Army Research Laboratory Human Research and Engineering Directorate Aberdeen Proving Ground, MD 21005-5425	**10. SPONSOR/MONITOR'S ACRONYM(S)** ARL/HRED
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

With new technologies for augmenting human capabilities developing at a rapid pace, problems with the current state of assessments are starting to emerge. Currently, assessments that are required to evaluate the effects and efficiency of these new technologies are often conducted unsystematically and in silo. In addition, theories have not kept pace with technological advancements and have not been useful in guiding research. Project Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness seeks to address these issues by revisiting the way assessments have been planned, executed, and used. We will also be looking into how assessment results can guide future research. We propose an approach comprising 3 phases. Phase I involves understanding various assessment needs, processes, and challenges within a domain in order to design a Web-based tool/product to help systematize assessment planning and execution. Phase II consists of developing a method for linking disparate research, while Phase III entails extrapolating the tools/products and linking approach to another research domain.

15. SUBJECT TERMS

current and future assessment challenges, standardized and systematic assessments, making sense of disparate research, linking approach, organizational capability

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Clayton Burford
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	SAR	120	**19b. TELEPHONE NUMBER (Include area code)**
Unclassified	Unclassified	Unclassified			407-208-3022

Contents

List of Figures

List of Tables

INTENTIONALLY LEFT BLANK.

1. Introduction

The current landscape of the US Army is complex, as it is in a state of transition, recovering from a decade and a half of sustained deployments. The Army is still engaging in meaningful action, supporting combat and humanitarian activities across the globe. However, it has the opportunity to prepare for the next enemy, developing Soldiers and units to respond to emergent missions for a multitude of future threats, many of which are distributed, chaotic, and perhaps overmatched (DA 2014a). Regarding the latter, the Army has a strategy to meet these demands for 2025 and beyond. *The Army 2025 Vision* (DA 2015) describes an agile organization that is supported by innovations in training, technology, and materiel to optimize human performance. The Army's differential advantage over its enemies is derived in part from the integration of these technologies with skilled Soldiers and well-trained teams (DA 2014b).

The Human Dimension Concept (DA 2014b) describes a vision for Army human performance optimization to develop Force 2025 and Beyond. That document identifies the importance of technology and human capital in optimizing human performance during strategy uncertainty. Particular emphasis is placed on the development of cognitive, perceptual, and physical capabilities in Soldiers and operational teams (human–human and mixed). Aligning with these strategic documents, the US Army Research Laboratory (ARL) *Human Sciences Campaign Plan* (ARL 2014) has focused on gaining a greater understanding of Soldier cognitive, perceptual, and physical performance. Augmentation research and development (R&D) efforts have focused on research to technologically enhance cognitive, perceptual, and physical aspects of human performance, while training R&D efforts have focused on achieving significant advances in Soldier training. The Human Sciences Campaign concentrates on high-risk, high-payoff transformational basic research, as well as technological innovations that generate unprecedented capabilities for future warfighters.

There have been significant successes in improving Army capabilities through both training and augmentation of the Solider and Soldier teams, but a consistent, uniform, and systematic approach to assessment of many current and emerging interventions and their effect on a given Army capability is lacking. This is a problem for the following reasons.

First, assessments cannot reliably, validly, or sufficiently be compared when assessments of interventions are not systematic. For example, Interventions A and B both claim to improve workload. Assessment of Intervention A uses Method A (e.g., informal interview after the task only) and Measure A (e.g., subjective workload measure), while assessment of Intervention B uses Method B (e.g., experimentation: pre- and postintervention) and Measure B (e.g., physiological workload measure). Although both assessments may yield results that suggest that their respective interventions help reduce workload and thereby improve performance, it is difficult to compare their results, as the assessments used different measures and methods. The measures may also not be suitable for the context of the assessment.

Second, the wrong or inadequate measures may be used when assessments do not follow a systematic or standardized process. O'Donnell and Eggemeier's (1986) psychometric criteria for measures provide one foundation for informing workload assessment decisions.

Third, most assessment efforts are unidimensional. Assessments should cover long-term effects of the interventions as well as the effects on other constructs. For instance, an intervention may reduce workload but increase complacency or have a more global impact on the unit. An intervention to help night vision might measure vision but not the effects on mobility or team situational awareness. There is a need for multidimensional assessments that show the relationships among constructs. These relationships should not be restricted to abstract relationships, such as those depicted by a nomological network, but also to "practical" relationships. For instance, whenever the intervention entails wearing something above a certain weight relative to the bearer's body weight and strength, the researcher should also assess mobility and physical dexterity.

Fourth, designers and consumers of assessment do not have an easily accessible means to know the "goodness" or "fit" of assessment options for their area of interest.

Fifth, assessments of technological innovations are typically focused on the effects of the particular innovation of interest, not on the combined effects of that innovation together with other technology that operators are already working with.

Last, much research is being executed in a vacuum and often researchers are unaware of other assessment opportunities. These issues occur for several reasons, including funding, the length of the publication cycle, the ability of theory to keep up with operational demands, and technology and mathematics outpacing theory and literature. Furthermore, issues reflect not just the decisions of individual research teams, but organizational limitations in supporting Army-wide standards for performing, evaluating, and communicating assessments.

2. UMMPIREE Overview

The Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness (UMMPIREE) project seeks to create an Army capability by revolutionizing the way in which research is conducted today, with particular focus on assessment. This technical report documents activities conducted during the first phase of the 3-phase UMMPIREE project. UMMPIREE comes under the direction of the Advanced Training and Simulation Division of the US Army Research Development and Engineering Command’s Human Research and Engineering Directorate.

The UMMPIREE project addresses the challenges of assessment applied to novel and emerging technologies. There are 2 critical issues with the current state of assessment. First, a consistent, uniform, and systematic approach to assessment of emerging interventions and their effect on a given Army capability is currently lacking. This issue is problematic, as the Army will need to make informed choices regarding the adoption of specific training and augmentation solutions. There is a great deal of variability in the use of current methods by which interventions are assessed, making it difficult to compare studies of similar types of interventions. There are many reasons for this. First, a multitude of tools and methods are available to conduct assessments. Assessments in research done by various laboratories, sites, and services are fragmented, with different researchers assessing the same constructs, such as workload, but using different measures, tasks, testbeds, study designs, and other methods and tools. While these reflect the variety of tasks encountered in the operations arena and different applications, unfortunately they also reflect the limitations encountered in research (e.g., funding constraints and inaccessibility of actual operators, which may require tasks to be simplified and measures to be modified for the student participant pool). Further, systems development (i.e., development of systems in support of Army capabilities) have often been stove-piped, and therefore the training and augmentation research in support of such systems development has tended to be stove-piped as well. Additionally, some successful measurement or assessment techniques, developed by individual research teams for a particular domain, have not been appropriately applied to, or shared across, other domains. Finally, individuals conducting assessments do not always have formal training as to how to follow best practices with regard to assessment activities.

The second critical issue is that existing assessment methods and tools need to evolve to meet future assessment needs and growing capabilities. There are 2 compelling reasons as to why this is the case. First, emerging technologies challenge the status quo of extant assessment methods and tools. Consider the example of human–robot teams (Hancock et al. 2011). There are currently no good measures of coordination for those teams. In addition, when technology opens up new research areas (e.g., human–robot teaming), there are often no strong theories to guide assessments and research direction (see Hancock et al. 2011). Researchers are more likely to conduct atheoretical research that only speaks to specific tasks and applications. These, in consequence, are less likely to generalize. This makes it difficult to build up a body of knowledge in the area, as conducted research tends to be piecemeal and expedient. The outcome is research that is unlikely to coalesce into a coherent body of knowledge, often reducing the likelihood of collaboration and exchange. Second, the world of assessment is faced with a growing series of advancements in real-time assessment methods of Soldiers' capacities in both operational and training environments. For example, advances in body-worn devices have enabled tracking of physical and cognitive states continuously and unobtrusively. Advances in network-based sensing have provided insights into patterns of team activities. These emerging capabilities require the development of new methods and tools by which assessments are conducted.

The remainder of this section reviews the current state of the art in assessment, together with the unique features associated with the Army context. Aims for UMMPIREE are guided by an understanding of the aims and purposes of assessment in the military, psychometric principles, and an appreciation of various challenges for the Army.

2.1 Current Assessment Principles and Army Application

2.1.1 The Army Context

The Army relies heavily on the results of human performance assessments to ensure force readiness (Hawley 2007). The Department of Defense (DOD) has made evaluation standards readily available (DA 2004; DA 2011) and there are published handbooks (Charlton and O'Brien 2001; Boldovici et al. 2002) that describe specific procedures for conducting training evaluation and human factors testing to include the design, development, testing, reporting, and/or reuse of assessments. However, these are primarily high-level guidelines. As a result, individual assessment teams, which may vary greatly in terms of their experience conducting assessment-related activities, must make a number of decisions often without sufficient guidance. Further, since there is a plethora of strategies, techniques, tools, and instruments that can potentially be applied, even experienced assessors may differ significantly in their choice of methods. This leads to a great deal of variability in methods, impacting the both the quality and comparability of the resulting assessments.

Therefore, while the assessment principles serve to guide the practice of assessments, there are often other factors and challenges impacting how assessments are conducted in the real world that are encountered throughout the various stages of assessments (i.e., planning, executing, and analyzing). In addition, the nature of military operations in the Army creates challenges for assessment. These include the dependence of outcomes on teams of varying composition, rapidly changing technology and training, and the need to conduct assessment under unstructured field conditions. Researchers often consider field conditions to be unstructured because in the field there are often many variables that can affect assessment outcomes that are out of the researchers' control. For instance, the ability to flexibly compose teams, while advantageous in dynamic military operations, can yield results that are not readily comparable. It would be difficult to assess the effects of an anticollision system of an aircraft on pilot performance when the data are from changing teams of pilot and copilots with varying levels of flying expertise, or when pilots could be undergoing new training during the assessment period, which typically stretches for weeks or months. In addition, field exercises can be prolonged or cut short due to the weather, and assessments are often conducted in conjunction with other activities which may invariably affect assessment outcomes. In contrast, in structured laboratory conditions where the research and assessments are the focus, such confounding and extraneous factors are better controlled and their influence on assessment outcomes minimized.

2.1.2 Definitions of Assessment

It is important to establish an initial definition and description of assessment. For the purpose of UMMPIREE, we define assessment as *an evaluation of the functional state, behaviors, or performance effectiveness of an agent in a given context at a given point in time, in terms of prescribed metrics*. Such evaluations are typically qualitative statements about the capability or readiness of the agent (e.g., passing a test, mastering a skill, or being ready for deployment) that are justified by quantitative data. The agent can be a person, a human-only or human–machine team, or a team of teams. The target operational context is the setting in which the agent takes, or fails to take, various mission-relevant actions. Generally speaking, the assessment context should possess functional fidelity for the target context (Sanders 1991); the assessment environment should elicit similar behaviors to that of the real operational environment. Functional fidelity typically requires that the complexity of assessment environment is scaled to that of the target environment, but it may not be necessary to reproduce all operational attributes to maintain the validity of the assessment.

An assessment is generated by collecting data about the agent and calculating a summary score quantifying that agent's performance. The score is compared with established performance standards or benchmarks. For example, the Army Physical Fitness Test (APFT) assesses each Soldier's muscular strength and cardiovascular endurance using 3 tests: push-ups, sit-ups, and a 2-mile run. For each test, the Soldier receives a score rated on a 0–100 point scale. Individual test scores are computed by comparing the Soldier's performance with age- and gender-specific standards. For example, to receive a minimally acceptable score (60%) on the push-up test, an 18-year-old male would need to perform 42 push-ups in 2 min; to receive the maximum score, that same Solder would need to perform 71 push-ups in 2 min. Similar scoring methods are used for the sit-ups and 2-mile run. Since a minimum score of 60 is required to pass each test, each Soldier's total APFT score can range from 180 to 300.

Such assessment data can derive from a variety of sources, such as self-reports, expert observer ratings, tools and technologies (e.g., simulators, radio networks), and sensors in the environment. Summary scores can then be computed using logical or mathematical formulae that code, combine, transform, and synchronize the raw data to create a summary about the agent's collective performance (Freeman et al. 2009). Formulae may be ad hoc or derived from an explicit statistical model. Scores may be quantified on a continuous scale (number of push-ups, percentage of correct responses on a test, etc.) or categorical (e.g., a 4-point subjective "readiness" rating). The summary scores are then used to make an evaluation of the agent in relation to some operational goal. This assessment may then inform decisions about the agent. For example, a Soldier who performs poorly on the APFT would likely be assigned some form of remedial physical training.

2.1.3 Purposes for Assessment

There are assessments conducted in support of performance improvements and assessments that are conducted in support of system, model, or theory development (see Fig. 1). The purpose of the former is to understand the impact of a given intervention (training and/or human augmentation) on human and mixed-team states, processes, and performance. For example, Soldier performance might be assessed before and after exposure to a particular training environment. Likewise, we could assess performance change in Soldiers using a particular augmentation device or concept. The impact of detrimental environments on performance may be assessed to specify effects on overload, stress and fatigue. The emphasis in these situations may be on either the individual Soldier or a mixed team. Assessments that are conducted in support of system, model, or theory development can be thought of as assessments with indirect aims that are focused on longer-term goals rather than the immediate impacts of particular operational environments or augmentation capabilities. Interpreting and deriving practical benefits from data often requires a theory of agent performance. Assessment may also be conducted in support of studies designed to test and develop theories of cognition, information processing, and operator state. Such studies may include developmental activities in which the researcher, developer, or experimentalist desires to understand the state of the Soldier or mixed team interacting with a particular system under development. Assessment is useful for defining and tracking the attributes of different Army cohorts, such as the "base state" of new recruits prior to training or the competencies of different operational specializations.

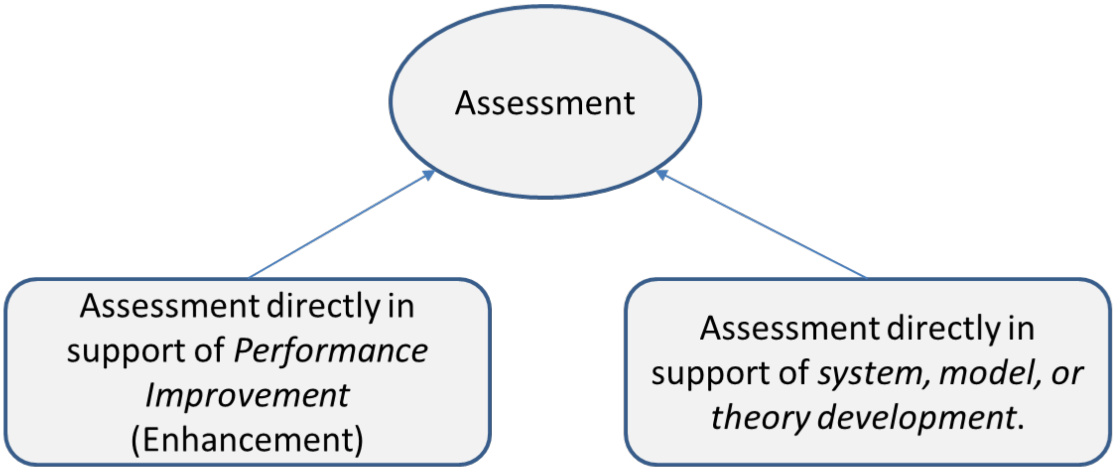

Assessment
Assessment directly in support of *Performance Improvement* (Enhancement)
Assessment directly in support of *system, model, or theory development.*

Fig. 1 Purpose of assessment

Although assessments have various purposes, UMMPIREE especially supports the assessment of the responses of humans and/or mixed teams to "interventions" intended to enhance performance, often taking the form of augmentation and/or training. The UMMPIREE project views this set of applications as an increasingly important need and is a significant element of the project. As technological capability continues to grow at an extremely rapid pace, we anticipate that many new training and augmentation capabilities will become available to the Soldier, the system developer, and human performance assessors. These future training and augmentation capabilities will almost certainly have unexpected consequences, both positive and negative, for the human and mixed-team. The mixed team itself is a complex concept that requires rigorous and novel ways of assessment.

UMMPIREE addresses the effects of inserting an augmentation and/or a training capability where or when one did not previously exist. This idea is shown in Fig. 2

and compares 2 situations, 1 without intervention and 1 with intervention. The bottom layer in the drawing represents the human dimension of physical, cognitive, and social characteristics. For given constructs, measurements are taken to assess the human (or mixed-team) performance while executing a given capability. The Training and Doctrine Command (TRADOC) definition of capability is used here. A capability is the "ability to achieve a desired effect under specified standards and conditions through a combination of means and ways across Doctrine, Training, Materiel, Leadership and Education, Personnel, Facilities to perform a set of tasks to execute a specified course of action" (DODD, 2008). Finally, the assessment is accomplished given the context of the Army capability or capabilities under consideration. It is at this level of assessment capability that UMMPIREE's initiatives will target.

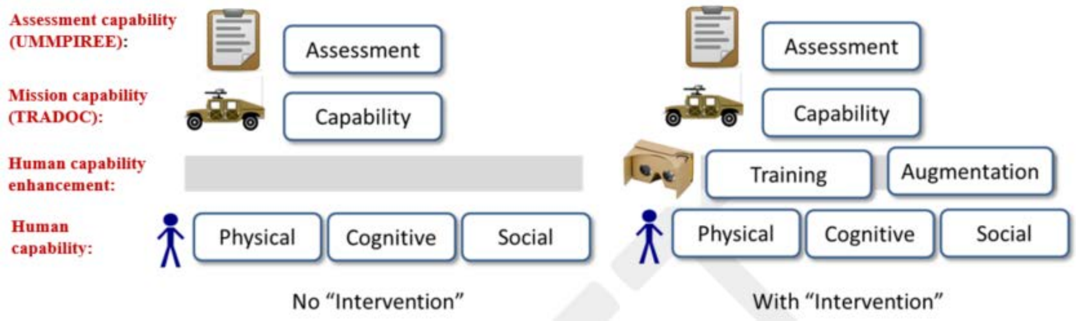


Fig. 2 UMMPIREE human dimension intervention-capability-assessment layers

Improving the validity and practical utility of assessment requires understanding of what should be done and what is actually being done. The ultimate aim is to define best practices for assessment that can be applied across domains, ensuring both validity and cross-domain compatibility of assessment. However, it is important to also retain the capacity to leverage existing findings, even when assessment has been unsystematic, methodologically flawed, or idiosyncratic to a particular domain.

2.1.4 Psychometric Contributions

Attaching meaning to numerical data obtained from measurement scales is central to assessment. The science of psychological measurement, or psychometrics, provides a basis for doing so, though it has some limitations in applied contexts. Psychological constructs such as skills or workload are not directly observable, and so psychometrics addresses *constructs*, hypothetical underlying characteristics assumed to influence observed behaviors. A construct is an abstract concept used in a particular theoretical manner to relate different behaviors according to their underlying features or causes (Heiman 2002). Constructs of interest to the military fall traditionally into the physical, cognitive, and social categories, although emotional and motivational constructs are increasingly recognized.

Classical psychometrics provided much of the early impetus for psychological assessment (Nunnally and Bernstein 1994). It assumes a simple correspondence between tests and underlying constructs (e.g., that an intelligence test is a direct measure of general cognitive ability). The observed test score is then reflects the sum of a notional "true score" plus some error. *Modern* test theory, exemplified by approaches such as item–response theory, Rasch scaling, and latent factor modeling, focuses on the statistical challenges of estimating latent constructs or traits from measured variables. Modern methods are diverse but have in common that they specify explicit statistical methods for estimating latent traits, requiring closer attention to item responses, error variance, and estimation methods, than is typical in classical psychometrics. In the military context, much psychological testing reflects the classical perspective that informed development of tests such as the Armed Services Vocational Aptitude Battery (Segall 2004), but modern approaches are becoming more prevalent.

It is central to assessment that measurements of the person or system are valid (i.e., that interpretations of quantitative scale data can be justified). Conceptions of validity have developed in parallel with the increasing statistical sophistication of psychometrics (Goodwin and Leach 2003; Plake and Wise 2014). Classical psychometrics encouraged validity as a static property of the test (e.g., the extent to which Stanford–Binet intelligence test scores could be interpreted as reflecting general intelligence). Contemporary testing standards, codified in the fifth revision of the *Standards for Educational and Psychological Testing* (AERA et al. 2014) define validity as the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Unpacking this definition, it sees validation as a process that involves building, testing, and refining arguments for test uses. It also recognizes the contextualized nature of validity in that it is tied to specific uses, requiring that real-world impacts of testing are recognized. The *Standards* also distinguish the following types of evidence that can be used to build a case for validity for some specified purpose:

- *Evidence based on test content.* Does item content correspond to the construct measured as evaluated by logical analysis or expert evaluation?

- *Evidence based on response processes*. Are responses on the test congruent with the construct, or do they reflect extraneous factors such as social desirability?

- *Evidence based on internal structure*. Are relationships between internal components of the test consistent with construct definition, as revealed by techniques including item analysis and factor analysis?
- *Evidence based on relations to other variables*. Are test scores associated with external variables consistent with construct definition and theory, as revealed by experimental and correlational studies?
- *Evidence based on consequences of testing.* Is there evidence that benefits presumed to accrue from test use can be confirmed?

Taken together, application of the *Standards* provides a road map for developing a scientifically (and legally) defensible case for using a test for a specified purpose. However, they present some challenges, such as the following, in the present context.

- *Reflective and formative constructs*. Psychometrics typically assumes that latent traits have a causal influence on measured variables. For example, in a confirmatory factor analysis, the model might be that scores on tests of hardiness, grit, adaptive coping, and the like reflect a latent trait for resilience influences. An alternative model is formative (Edward and Bagozzi 2000), in which the construct reflects the outcome of multiple factors that may not be correlated. For example, we could assess “fitness for duty” from indicators such as hours of sleep, blood alcohol concentration, and cardiovascular functioning. The indicators do not define any latent trait or construct, but they could be used to derive a practically useful index, though one requiring different validity arguments to those justifying a typical reflective measure. As acknowledged by Schmittmann et al. (2013), the challenges associated with establishing or established causal relations strains the validity of reflective and formative models. Regardless, modelers often attempt to “best fit” causal relations within reflective or formative models. These model approaches have significant causal relationship problems, namely:
    - Role of time. They do not sufficiently address or capture the dynamic and cyclic nature of variables.
    - Inability to articulate processes. The causal relations between constructs and indicators are poorly characterized.
    - Relations between indicators. The relations between the indicators or constituents are less important than the latent variable.

  These problems may well be addressed by a network approach (Schmittman et al. 2013) that UMMPIREE will be examining. The network approach accommodates these issues naturally, and in a way that no reflective or formative model can do because it allows us to reason about dynamics within the psychometric context of the indicator variables themselves. The essence of a network construct is not a common cause; rather, it resides in the relations between its constituents. Network structure and node properties and dynamics may also be useful in expressing individual differences in relation to the construct (Schmittman et al. 2013). In addition, studies investigating the relation between network properties (e.g., distribution of connection weights or the number and type of equilibrium points) on one hand, and the possible range of configurations of cross-sectional data obtained at a single time point (e.g., data conforming to a 1-factor model or a 5-factor model with correlated factors) on the other may provide useful starting points for dynamic accounts of psychological constructs (Schmittman et al. 2013).

- *Assessment beyond the individual*. The *Standards* focus on measurement of constructs in individuals, in keeping with their origins in psychological and educational research (AERA et al. 2014). However, UMMPIREE also emphasizes assessment of teams, technology-based systems, and training procedures, going beyond the focus of the *Standards*. To some extent, the challenge can be met by insightful extension of the same principles. We may treat the team rather than the individual as the unit of assessment and look for the same forms of evidence to support the validity of a measure of team effectiveness. The relationship between measures of individual and team functioning remains a challenge (Funke et al. 2012). Traditional technology lends itself to a “classical” perspective (e.g., in measuring the bandwith of a communication channel, for which there is a meaningful “true score”). The behavior of autonomous systems, however, increasingly emerges in complex and unpredictable ways from the hardware and software, and measurement is thus increasingly directed toward unobservable constructs, as in humans. Assessment of training methods may be reduced to measurements of impacts on individuals, although questions may remain about training x aptitude interactions (Snow and Swanson 1992) (i.e., whether impacts differ across individuals).

- *Technological and cultural change*. The *Standards* tend to assume a constant world in which presumably universal constructs such as cognitive ability and physical fitness generalize across cultures and across generations. Military technology and the uses to which it is put change rapidly, so validity arguments may have a very limited shelf life. Under the impact of a changing technological culture, the skills and attitudes of military recruits are also likely to change rapidly. A validity argument sound for an earlier generation may not hold up for digital natives.

- *Practical constraints*. Psychometrics assumes that data can be collected from large samples in controlled conditions. However, military assessment data might not meet these requirements. It may also be important to integrate evaluations from experts in the field into assessments. Psychometric purists would discard data collected from, say, a small number of Soldiers operating in combat conditions, but such data add a dimension to assessment that may not be recoverable in any other way.

2.2 Meeting Assessment Challenges: Research Needs for the Army

The previous section made the case that measurement of constructs must support an evidence-based justification for using assessments for specific purposes; that is, establishing validity is critical for assessment. It also laid out some of the assessment issues that are common across a variety of contemporary settings. Assessment is especially challenging in the military context because of the complexity of the problem along multiple dimensions. The targets or objects of assessment are diverse, as are the constructs used for assessment. Existing practices are varied and sometimes unstandardized, reflecting the different purposes and contexts that drive assessments, as well as needs to assess in field environments as well as in controlled settings. The following challenges define research needs to which the UMMPIREE project should contribute.

2.2.1 Challenge 1. Constructs Are Numerous and Diverse

Fundamentally, UMMPIREE is focused on activities involved with assessing the impact of intervention on the human dimension (Chen et al. 2014), often in the context of interaction with technology. Even a single intervention may produce a diverse range of outcomes, including impacts on the physical, cognitive, and social aspects of the individual and teams. For example, augmenting one or more team members with an exoskeleton obviously enhances physical capabilities. However, the intervention may also improve cognitive function (less distraction from physical exertion) and social relationships (other team members may be admiring or envious). A full assessment may require measurement of multiple constructs. Indeed, the term assessment is used instead of “measurement” to indicate that assessment is typically a complex process involving multiple measurements from multiple dimensions.

2.2.2 Challenge 2. Variability in How Assessments Are Conducted

Often there may be several qualitatively distinct methods for measurement of a construct that do not necessarily converge. For example, although workload is an intuitively accessible construct, it may be measured via self-report, psychophysiology, or behavioral indices. Each measure tells the assessor something different about the operator's response to task demands, and each has its own strengths and limitations (Matthews and Reinerman-Jones, in press). It is thus necessary to develop frameworks for systematic evaluation of candidate measures for an assessment that balance breadth and depth of coverage of constructs with practical considerations.

UMMPIREE is ultimately interested in providing tools that help standardize the "how" of assessment. Ultimately, we seek to achieve this aim by developing a more comprehensive understanding of assessment processes, procedures, and methods. We approach the "how" of assessment from a broad as well as targeted perspective. From a broad perspective, we refer to the overall activity of conducting an assessment. This include activities to design the assessment plan, select the methods (tools, participants, context, measures, etc.) that will be used, conduct the analyses, and report the results. From a targeted perspective we look to focus on specific areas of conducting an assessment. As one example, specific areas of great variability are 1) the conceptual definition of a construct and 2) the selection of measures to address a given construct. The idea of measurement represents at least a portion of the challenge of UMMPIREE in that a given measure might be—appropriately or inappropriately—defined and used differently in different contexts. In that sense, one of the goals of UMMPIREE is to help users (e.g., the researcher or system developer) identify the best measures for the given assessment use case. The goal of UMMPIREE is also to standardize these measures as much as possible so assessment results across domains and use cases have the desired uniformity and reliability.

2.2.3 Challenge 3. Assessments Cover Both Humans and Machines

Assessment, in the UMMPIREE context, refers to individual humans and human teams, as well as such humans and human teams paired with intelligent, autonomous, or semi-autonomous machines. Developing a common assessment framework for humans and machines is novel but essential for implementing the Third Offset Strategy, which envisages seamless human–machine teaming in applications including human–machine collaborative decision making, augmented human operations, and advanced manned–unmanned systems operations (Ellman et al. 2017).

The UMMPIREE project draws a distinction, although not a completely impermeable one, between systems, however complex, that have limited intelligence capability (such as a traditional command and control system or a shoulder-fired missile) and machines that have a significant ability to behave with some intelligent autonomy such as robot or some unmanned vehicles (see Hancock 2017). This distinction also draws a contrast between assessment that is viewed in UMMPIREE as “human-centric” and test and evaluation (T&E) that is viewed as “system-centric”. In both cases, assessment and T&E, there is interest in the human state, behavior, and performance, but considerably more emphasis is placed on the human in assessment than T&E. The term “mixed agent” is used to connote a teaming of humans and (intelligent) machines. The UMMPIREE scope does not include assessment of machine intelligence without a human partner or teams of such quasi-independent intelligent machines.

UMMPIREE’s scope includes Soldiers in various military occupational specialties and ranks who are learning, training, or participating in some operational task in support of an experiment (e.g., combat development) or system development activity. There is also the potential for other nonmilitary (e.g., civilian) use. A long-term goal also includes Point of Need (PoN) and continuous assessment (National Council on Measurement in Education 2017), which could extend the types of activities to real-time operational actions in which the Solider or mixed team is engaged while under assessment.

2.2.4 Challenge 4. Multiple Contexts and Timespans for Assessment

Assessment may take place at different stages of a Soldier’s or civilian’s career and include assessment of knowledge, skills, abilities and other characteristics (Krumm and Hertel 2013) from novice to expert. Purposes of assessment vary in that the aim may be to determine Solder aptitude for a specialization, outcomes of learning or training, or to support system development or model and theory development. In a development environment, the developer is interested in how the system of concern interacts with the Soldier and/or the mixed team. This type of assessment could occur early in a system acquisition in prototype phase, but could also occur in an operational test and evaluation phase late in the system acquisition process. The challenge is that the relevance of constructs may vary according to context. For example, trainees’ workload might be assessed to see that they were not overloaded with novel information. In experienced operators, the concern may be to identify specific task elements that are overly demanding. Can the same workload assessment be used to answer these different questions? The assessment process needs to incorporate some means for determining the most-suitable measures to address different purposes. Similarly, the temporal focus of assessments differs.

Measurement of stable, dispositional Soldier characteristics such as cognitive ability and resilience may inform assessment across all career stages. However, assessment of an acute stress response might be relevant only to the specific environment evoking the response. Again, a systematic approach to determining the timespan for assessments is needed.

2.2.5 Challenge 5. Variability in Assessment Locations

Traditional psychometrics assumes that tests can be administered in controlled, standard conditions, an assumption that conflicts with the military context in the real world, where assessments occur in different locations including classrooms, laboratories, real and virtual training environments, and operational settings in which the assessment may not be the priority activity, and where the influence of extraneous factors is not easily controlled. It is expected that UMMPIREE will support assessment in such a variety of assessment locations. An important location for future UMMPIREE capabilities is the PoN, which could include home stations and deployed locations as well as more-conventional training locations. This includes the concept of "real-world" assessment; that is, assessment in an operational context.

3. UMMPIREE Goals and Objectives

The overarching goal of UMMPIREE is to contribute to building an organizational capability for research in human capability enhancement by focusing on assessments, which are critical for research, application, and organizational effectiveness. UMMPIREE seeks to improve assessments with regard to existing methods and practices, as well as future methods and practices with novel technologies. Organization-wide initiatives typically require the right personnel, processes, and technology to be in place. The tools and knowledge products developed from this project seek to help equip assessment personnel, improve assessment processes, and put in place technological support necessary for quality assessments.

Central to the UMMPIREE effort is standardizing and unifying assessments. Assessments in applied research that must accommodate limitations in resources, contextual constraints, and serve multiple agendas should still be able to withstand scientific scrutiny. Systematization and standardization are necessary but not sufficient conditions for quality assessments. However, assessments that are scientifically defensible would have been systematically planned and executed in a standardized manner to ensure validity. Providing the appropriate knowledge and tools to guide the systematization and standardization will increase the likelihood of quality assessments. UMMPIREE seeks to enhance the quality and impact of assessments by formulating tools and products that facilitate systematic planning and execution of assessments that are standardized. Products from UMMPIREE can also help identify the skills required to execute quality assessments, facilitating matching of skill sets for formation of assessment teams.

Improving assessments at an organizational level entails more than helping assessment teams execute systematic and standardized assessments. UMMPIREE seeks to unify assessments across the board and impact the way researchers determine the type and number of assessments to be conducted. As a result of many of the challenges described in Section 2.3, researchers and assessment teams have not been able to leverage as much from the assessments conducted across labs and services. This can lead to some degree of effort duplication where research does not yield the utility that is commensurate with the resources expended. UMMPIREE anticipates developing tools and knowledge products that will help assessment teams discover overlaps in research and link research that may seem disparate at first glance, encouraging collaboration. These tools should also help various assessment stakeholders (e.g., researchers, engineers, managers) make sense of, and extract from a body of research, information that help meet their assessment needs.

3.1 UMMPIREE 3-Phase Plan

As part of this first year, the UMMPIREE team developed a 3-phase plan by which the UMMPIREE goals are to be accomplished. Each phase is made up of several activities, each of which addresses a specific objective. That 3-phase plan is discussed in this section, with the remainder of the report focusing in more detail on the activities conducted as part of Phase I.

3.1.1 Phase I: Baseline

The broad objective of the first, baselining phase was to 1) establish the current state of assessment practices for determining UMMPIREE's direction and tools/products to develop, 2) determine subsequent project activities, and 3) provide a point of comparison against future state of assessments when evaluating UMMPIREE's impact. Separate activities aimed to

- specify terminology, construct definition, and assessment principles,
- develop a conceptual framework and architecture for characterizing and guiding assessment,
- identify a domain that exemplifies assessment challenges and needs, and

- establish a baseline of current assessment needs and practices in the domain that can inform ideas for tool and product development.

The purpose of Phase I was to define the problem and scope the efforts for Phases II and III. That is, the baselining phase progressed from understanding normative and descriptive perspectives on current assessments to developing products and tools in a practical and grounded way, transiting the project to the next phase. This section of the report describes the activities constituting the Phase I plan while other sections will describe the Phase I products in more detail.

Activity 1: Developed tools to promote a common language and mental model/representation that can help communicate assessment ideas across different users of assessments. As part of this activity, we developed the following products:

- Outline of the principles of assessment. Describes the principles guiding Army technology assessment.

- Assessment lexicon. A list of common terms and definitions used in assessment.

- Conceptual Assessment Model (CAM). Describes an approach to categorize and describe conceptual and operational definitions of constructs as well as their measures.

Activity 2: Developed a conceptual architecture to map assessment processes and activities including planning, execution, reporting, analysis/postexecution after-action review (AAR), and data fusion. The activity resulted in the following products:

- Assessment Planning Framework (APF). Describes the current process by which a quality assessment is conducted.

- Mobile Assessment Architecture (MAA). Presents a high-level functional architecture by which to organize future assessment solutions.

Activity 3: As part of Phase I, the UMMPIREE team selected the research domain of human–agent/–machine/–robot teaming (HMT) to begin our focused effort in Phases II and III. This domain exemplifies issues and challenges with current assessments (e.g., rapid pace of innovations and technological advancements plus a lack of strong theories in domain). It is also believed to be a domain that would enable the Army to reap the most gains from the UMMPIREE project. The activity resulted in the following:

- Target domain identified. The domain enabled subsequent effort to be based on real-world examples and cases.

Activity 4: Once the HMT domain was selected, the team then explored the current state of assessments and engaged researchers in that domain. The purpose was to identify a baseline against which to compare future assessments. This effort aimed to provide real-world data to ground project efforts and to engage assessment stakeholders and obtain their insight and collaboration for the project. It also helped identify important factors that shape research and assessments (e.g., constructs, measures, domains, environments, practices, and drivers of research). The following products were developed during this activity:

- Baseline of current research practices and assessment needs. Collected from HMT subject matter experts (SMEs).
- Research Database Capture Tool. A data-collection tool.
- Research Visualization Tool. A first step toward finding a method to link research.

3.1.2 Phase II: Linking and the Network Approach

During Phase I, we built a plan for Phases II and III. Phase II will address the issues of assessments that relate to fragmented research in domains that do not look to strong theories to guide research, as well as the problems that have begun to surface from assessments not evolving along with technology and other innovations. In the former realm, UMMPIREE will develop a method to link assessments from different research studies to guide research systematically and build up a knowledge base. There are 2 parts to this: 1) deriving a conceptual basis for linking and 2) formulating a mathematical method for linking. Baseline data from Phase I will be used to test the linking approach. The team will also work on the technical infrastructure needed to support the effort. We envision a Web-based, Web-accessible product or application (app) with multiple modules and plug-ins that connect to a data store. The latter will include novel data sources that can identify and encourage the use of new measures and assessment methods. This will help assessment keep pace with evolving methods and technology. The activities in Phase II aim to achieve the following:

1) Derive the conceptual basis for integrating studies that employ different measures for the same constructs.
2) Formulate a mathematical approach for such linking. Approaches to consider include the network approach, social network analysis that can use some techniques from data mining like clustering, nearest neighbor methods, and the like.

3) Identify and develop prototypes of products/apps with the supporting technical infrastructure, for instance:

   i) A tool where users can be guided through a “wizard” to input their assessment needs and have as output an assessment plan and execution details.

   ii) A tool that captures users’ inputs on their assessment experience. This can then “feedback” into tool (i) to generate future assessment plans and execution details.

   iii) Create visualization of SME research to help to see “links” among their research. This can help with the exploration of methods to link up research.

Products expected to result from Phase II activities include the following:

- White paper on the concept approach to linking. Presents the conceptual effort and implications of mapping research across domains, constructs, and measures.
- Mathematical exploration of linking approach with application to HMT research. Development and evaluation of linking approach and derivation of a validation plan for the approach.
- Prototype of a Web-based, Web-accessible application. A tool to help planning and execution of assessments to be more systematic and standardized.
- Conceptualization of the technical infrastructure. A development plan for the technology required to support tools/products developed.

3.1.3 Phase III: Cross-Pollination

Phase III of the project involves extending the developing products of the project to other domains and other assessment stakeholders. The work in this phase will focus on further development and refinement of the products and tools so that they are more customizable and readily adaptable to various assessment contexts. The team will also identify another specific domain (likely to be related to the initial domain; e.g., human–robot interaction [Hancock et al. 2011]) to human–automation interaction (Schaefer et al. 2016) on which to apply the linking/network approach. Engaging these assessment users in the new domain will then become a priority. The activities in Phase III aim to achieve the following:

1) Extend the approach selected/developed to other research in other remote domains.

    i) For instance, from knowing about workload assessed by subjective measures in the military domain, network science might be applied to understand workload assessed with subjective measures in another domain, such as aviation. Workload and resilience, as researched in aviation, can allow us to test a new sensor technology such as electromyography and perhaps the generation of a new construct altogether termed, such as Workload Strain.

2) Enhance assessments across the board to facilitate inter-lab collaborations and technology transfer agreements.

Products expected to result from Phase III activities include the following:

- Identify new research domain. Identify another research domain facing similar assessment challenges to which UMMPIREE products and tools can be extended.
- Modification and extension of tools/products. Tools and products to improve planning and execution of assessments can be extended to improve assessments in other domains. Linking approach can be applied to help coalesce research in other domains.
- Extend project initiatives to other stakeholder tiers in the HMT domain.

3.2 Tools and Knowledge Products

UMMPIREE wants to build organizational capability in assessments. This entails not just improving the quality of assessments, but also improving research that heavily depends on the quality of assessments. Central to this goal, the tools and knowledge products from UMMPIREE should enable ready transition to selected research communities. The project team foresees that UMMPIREE tools and products would be particularly helpful for researchers in the training and augmenting technologies space. We also envision that the project deliverables would contribute to the training and readiness community in terms of helping them with gap analyses. Improving research with UMMPIREE products across the board includes the following:

- Specifying strategies that enable researchers to leverage each other's research despite the fact that they all use different methods, measures, tasks and environments. Enhanced awareness, communication, and synergy of complementary research efforts will improve the efforts themselves, strengthen their practical impacts, and minimize duplication of effort where redundancies exist. Tools should ultimately support Human Capabilities Enhancements (HCEs); technologies that directly or indirectly enhance the human's cognitive, perceptual, or physical competencies in support of the organizational mission.

- Providing tools for researchers to use to discover conceptual and methodological overlaps in their research. Discovery of commonalities in approach will support collaborations that enhance applicability of research across contexts, enhancing operational capabilities.
- Providing tools to match/assemble skill sets required for quality assessments across the services, again leading to HCEs, and prioritization of work that maximizes those enhancements.
- Providing tools for researchers, engineers, and managers to easily extract the actionable outcomes they need from a body of research in support of HCEs.
- Helping researchers developing training and augmentation interventions for HCE to assess combined effects of interventions across time and space. Tools should support choice of assessments to optimize products of validation studies for both the research lab and the organization.
- Helping researchers in the training and readiness space to perform gap analyses to understand strengths and limitations of current assessments, driving systematic enhancement of assessments and research practices.

3.3 Summary of Activities for Phases I and II

Table 1 summarizes how the activities support the goals and objectives of UMMPIREE. It lists the problem addressed by each activity, how the activity contributes to solving the problem, and the tools and products generated by the activity that support advancements in methodologies for HCEs. (Phase III is not included because activities will be shaped by Phase II outcomes.) Phase I and part of Phase II efforts are discussed in more depth in the sections that follow.

Table 1 Rationale for UMMPIREE activities

Initiative from activity	Why is activity important?	How activity addresses that problem	How activity feeds into tool or knowledge product
Phase I			
Activity 1: Specify terminology, construct definition, and assessment principles			
Assessment lexicon: A list of common *terms* and definitions used in assessment.	Current assessments tend to be unstandardized and unsystematic in the use of terms and language.	Assessment lexicon helps everyone have common understanding of the terms used in assessment.	Web-based tool to help new assessors can contain a look-up glossary of terms.
Conceptual assessment model (CAM): Describes an approach to categorize and describe *conceptual* and operational definitions of constructs as well as their measures.	Current assessments tend to be unstandardized and unsystematic in the use of terms and language. These arise because assessments often involve system and computer engineers, psychologists, etc.	The CAM described common ideas in assessments in terms of the Unified Modeling Language (UML), which is familiar to engineers. Describing how constructs are operationalized and the like in UML will help foster common understanding among psychologists and engineers.	Web-based tool can have the option to describe assessments in UML.
Principles of assessment: A *description* of the principles guiding Army technology assessment.	Current assessments tend to be unstandardized and unsystematic in practices and procedures because different people (even different psychologists) have dissimilar ideas of assessment principles.	If assessors adhere by a common set of assessment principles, assessments may be more systematically conducted (i.e., will have to first define the purpose of assessment before planning it).	The principles of assessment can be used to generate checkpoints in the assessment "wizard" that guides a new assessor through assessment planning.
Activity 2: Develop a conceptual framework and architecture for characterizing and guiding assessment			
Assessment planning framework (APF): Describes the current process by which a quality assessment is conducted.	Current assessments tend to be unstandardized and unsystematic in practices and procedures.	The APF outlines the factors affecting decisions as well as the best practices during the planning stage of assessment. It shows how to conduct the best assessments given real-world issues.	Web-based tool can walk the new assessors through assessment planning while highlighting points during the planning where decisions made impact assessment quality the most.
Mobile assessment architecture (MAA): Presents a high-level functional architecture by which to organize future assessment solutions.	Current assessments tend to be unstandardized and unsystematic in practices and procedures.	Unlike the APF, which focused on the planning stage of assessment, the MAA shows 5 stages of assessment. It outlines differences in involvement level with respect to different users. For the different users, it shows the system components that are relevant that user.	Web-based tool can have 3 versions for each of the 3 users. Each will take the user through what she/he is most concerned and interested in with regard to the stages of assessment. It will also enable users to see assessments from other user's perspective.

Table 1 Rationale for UMMPIREE activities (continued)

Initiative from activity	Why is activity important?	How activity addresses that problem	How activity feeds into tool or knowledge product
Activity 3: Identify a domain that exemplifies assessment challenges and needs			
Target domain	We needed a domain area to test out ideas for improving assessments. This area needs to show the main problems we find will assessments and will be most likely to benefit from the project.	Domain will help team to focus on understanding assessments in a real-world and appreciate the assessment issues highlighted (e.g., assessing competing innovations in HMT).	Web tool should be mocked-up to help assessments in HMT first. Contents within HMT such as commonly cited theories, constructs, and measures used can be used to generate recommendations for the new assessor.
Activity 4: Establish a baseline of current assessment needs and practices in the domain that can inform ideas for tool and product development			
SME interviews	Current assessments tend to be unstandardized and unsystematic, and we need to see what factors are shaping assessments in the real world that contribute to the unstandardization and unsystematization.	Helps us understand assessment practices and needs of actual researchers in HMT domain. Understand what factors make assessments unsystematic and unstandardized.	Meeting with and interviewing SMEs helps developers of tool to understand the most immediate needs of assessors in the real world. Tool development will be "grounded" to serve this HMT research community first.
Research database capture tool (RDCT): Same function as SME interviews, but compared with SME interviews this a knowledge elicitation method that demands fewer resources.	Current assessments tend to be unstandardized and unsystematic, and we need to see what factors are shaping assessments in the real world that contribute to the unstandardization and unsystematization.	Helps us understand assessment practices and needs of actual researchers in HMT domain, and what factors make assessments unsystematic and unstandardized.	Info from SMEs helps tool developers understand the most immediate needs of assessors in the real world. Tool development will be "grounded" to serve this HMT research community first.
Research Visualization Tool (RVT): A first step toward finding a method to link research. This activity transits the project from addressing the unstandardization and unsystematization of assessments to finding an approach to link research.	Current assessments tend to be unstandardized and unsystematic. The visualization can help show that there may be some linkages and structure that can be made to organize research (by construct or by author/collaborator, etc.).	Activity shows maps and networks of researchers and the constructs they study, as well as their collaborators.	This can be further developed to be a "plug-in"/module for the Web-based tool to help assessors visualize relationships among constructs and measures. Can also help management to visualize relationships among researchers, collaborators and constructs (e.g., see where research "hotspots" are).

Table 1 Rationale for UMMPIREE activities (continued)

Initiative from activity	Why is activity important?	How activity addresses that problem	How activity feeds into tool or knowledge product
PHASE II			
White paper on the conceptual basis for linking *Mathematical exploration of linking approach with application to HMT research* *Prototype of a Web-based, Web-accessible application* *Conceptualization of the technical infrastructure*	With the current state of assessments, we may not be suited to deal with new challenges that come with novel technologies and new methods. In areas without strong theory to link research (i.e., HMT), there is a need for another way to link research. We are working on a bottom-up network approach to do this.	Activities result in a network approach to link research.	Web-based tool can show links among similar research, and identify new measures and the like for the assessor based on linkages.

4. Phase I (Year 1): Baselining

4.1 Activity 1

For the first step in baselining, it was necessary to revisit the purpose and use of assessments in the Army context. As previously discussed, the scope and diversity of assessment procedures utilized by the military mitigates against standardization of methods and against ready communication between different stakeholders. First steps to managing this diversity include clarifying general principles underscoring assessment for military goals, developing a standard terminology for assessment concepts, and developing models for framing assessments within a common structure. Activity 1 was directed toward these goals.

4.1.1 Principles of Assessment

If assessments are to inform wider organizational decisions and serve larger organizational needs, they need to be executed with the right intent and purpose that fits in with that of the organization. When these have been ensured, steps can then be taken to systematize and standardize assessments. General psychometric standards are well-established, as discussed in Section 2.2.3, but principles for Army application require clarification.

4.1.1.1 Problem Statement: Need for Guiding Principles

As with any organizational initiative, conducting assessments should be aligned with the organization’s mission and vision. Assessments that are not compatible with larger organizational goals may be executed well from a scientific standpoint but may not serve broader purposes adequately or, worse, may send a message that runs counter to organizational values.

4.1.1.2 UMMPIREE Solution: Identify Key Principles of Assessment

The UMMPIREE team outlined the following 4 principles of assessment that should guide how assessments are executed in the Army. These may be incorporated as prompts in future tools that can be developed (see Appendix A for details).

1) **Assessment must be grounded in a clear purpose.**

   Assessments can be used for multiple purposes: selection, determining overall readiness, investigating training or program effectiveness, technology integration, human augmentation, changes to doctrine, or changes to tactics, techniques, and procedures. In addition, assessments can provide formative (e.g., to give feedback or guide further development) or summative feedback (e.g., to determine mastery or impact).

2) **Assessment must align with the Army's Mission, Values, and Warrior Ethos.**

   In addition to supporting mission needs, assessments should reflect and reinforce the Army's cultural values and core beliefs that transcend time and mission if they are anchored in Army Values and Warrior Ethos. The object of assessment must be directly traceable to key elements of the Army's Mission, Values, and Warrior Ethos. In addition, the process of assessment must be consistent with the Army's Mission and Values.

3) **Assessment should serve multiple users, uses, and time frames.**

   Since no one measure can provide complete information, and assessments are not one-size-fits-all needs, multidimensional and multimodal assessments are needed.

4) **Assessment must produce information whose value exceeds its cost.**

   To maximize their value, assessments should typically be conducted in a timely manner so that the information they yield is actionable. Furthermore, the resource costs from assessments should be minimized to maximize return on the investment in the assessment. Thus, assessments should not divert resources from actual training and operations. They should also be sustainable and feasible in the target environment.

4.1.2 Developing an Assessment Lexicon

Assessment teams typically comprise personnel from different disciplines. This is necessary because conducting assessments requires a varied skill set. For example, a typical assessment can involve a researcher defining research design, an engineer operating the technology being assessed, and an information technology or software engineer providing support for the simulated environment/tasks and data collection. However, the opportunity for miscommunication increases with interdisciplinary personnel who are trained differently. The notion of measuring abstract constructs (situational awareness, workload, etc.), while commonplace for social science researchers, may not be familiar to those from other disciplines such as engineering. Poor team communications can contribute to assessments being unsystematic and unstandardized.

4.1.2.1 Problem Statement: Lack of Shared Definitions of Assessment Terminology

Some of the problems in communication within assessment teams arise because the same terms can carry different meanings in different disciplines, or the meaning of certain words in the English language is different when used in the assessment context. For instance, to engineers, a "model" may be a representation of a system with its constituent parts, while to a psychologist or social scientist it may be a theory of how variables are interrelated (Stowers 2015). Compared with its general use, the word "validity" denotes much more in psychometrics, where there are notions of different types of validity. Miscommunications that stem from such differences in definition of widely used terms in assessment can affect how assessments are conducted.

4.1.2.2 UMMPIREE Solution: Compile a Lexicon of Commonly Used Assessment Terms

A lexicon of commonly used terms in assessments was created to facilitate common understanding of assessment terminology. Its compilation ranged from terms related to measurement such as "metric" and "construct validity" to terms associated with statistical analyses of data collected (e.g., "analysis of variance" and "confidence interval"). The lexicon can be incorporated into future tools and products to be developed to assist different members in the assessment team with various assessment stages (see Appendix B).

4.1.3 Conceptual and Operational Definitions of Construct

Constructs are a central component in assessments and, because they are often abstract, are not readily observable or quantifiable. Validation of constructs is a complex process that should be tied to specific goals or uses for the measurement of the construct (see Section 2.2.3). Hence, a central aspect of assessments is the operationalization of constructs through their indicators, or measures. Assessment teams need a shared understanding of this notion and its implications to be able to work in concert to execute quality assessments.

4.1.3.1 Problem Statement: Need for Common Language to Describe Assessment Concepts and Processes

In addition to shared terminology, the assessment team has to have shared understanding about the various components necessary for assessments and the processes in which these all come together. Communication of these is impeded in part due to the lack of a common language among multidisciplinary assessment teams.

4.1.3.2 UMMPIREE Solution: Express Critical Assessment Concepts and Processes in a Common Language

Describing the assessment components and processes using UML can help improve communications among assessment team members. UML is a general-purpose developmental modeling language in the field of software engineering that is intended to provide a standard way to visualize the design of a system, is commonly used in engineering, and is fairly easy to master. A UML diagram can express different concepts and components and how these all work together.

UMMPIREE developed the CAM to help those not familiar with the social sciences understand assessments conducted in the discipline (see Appendix C for details). The CAM used 2 concepts from UML, classes and compositions, to model the relationships among the core components of assessments. It includes the "Operationalized Construct", the "Mission/Assessment Context", and the "Subject Model", corresponding respectively to the construct operationalization and measure selection, the task and environment, and the subjects involved in the assessments (Fig. 3).

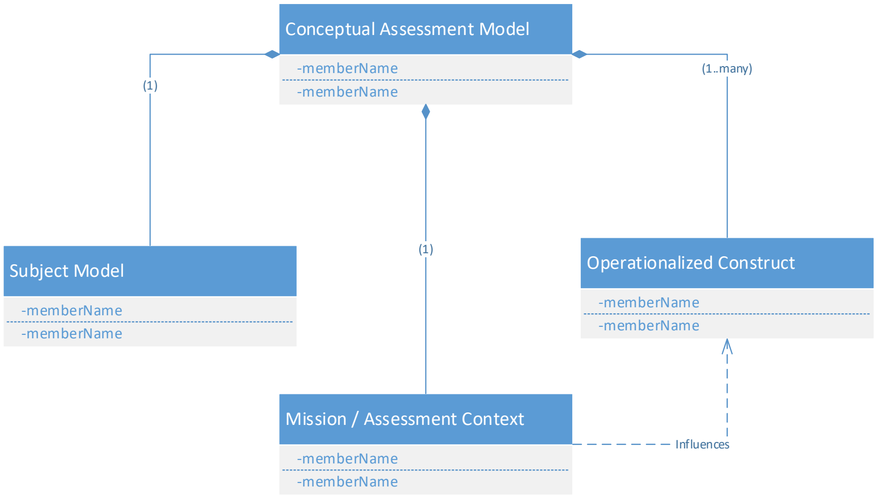


Fig. 3 Conceptual assessment model (CAM)

The Subject Model depicted different categories of the unit of analyses in the assessments and specified the "elements" in each category. For instance, an assessment could be conducted at 3 levels of analyses: human, robot, and the human–robot team. Correspondingly, the Mission/Assessment Context component indicated the context of the assessment with respect to the human, the robot, and the human–robot team. The Operationalized Construct component specified how the constructs have been operationalized given the Mission/Assessment Context. It also captured the basis for the operationalization in terms of the theory and empirical evidence in the literature supporting the operationalization.

In the example of an assessment conducted to assess trust in a human–robot team comprising a Soldier and a sniper robot, the Subject Model would indicate that the unit of analysis was human (i.e., trust of the Soldier), and the Mission/Assessment Context would state that assessment context was a simulated intelligence, surveillance, and reconnaissance mission. Under Operationalized Construct, the construct of trust would be shown to be operationalized by performance on a mission-specific decision, such as frequency at which Soldier accepts the recommendation of the sniper robot, and by specifying that this operationalization of trust had been based on previous studies on trust in a similar context.

In summary, the CAM helps to convey to all in the assessment team that assessments require them to

1) Identify and specify the components that will be used for a particular assessment.

2) Develop a data collection and measurement plan for each element of the CAM that is identified as useful for the assessment.

3) Articulate how the components and elements relate to one another from an analysis perspective (e.g., how do the tasks relate to the constructs? What data elements will be used for calculating what assessment measures?).

4) Articulate how these data will be analyzed and specify any other higher-order measures such as measures of performance and measures of effectiveness.

4.2 Activity 2

The quality of assessments depends heavily on the extent to which they preserve scientific rigor, despite practical constraints, and yield results that are useful to different stakeholders. Scientific rigor is often compromised by the use of unsystematic procedures and practices. One of the principal goals for UMMPIREE is to develop a general framework for conducting assessments systematically across diverse Army contexts. Activity 2 aimed to address this goal by identifying best practices in systematic planning for assessments and in accommodating the needs of multiple stakeholders.

4.2.1 Systematic Planning for Assessments

A major limitation of existing assessment practices is their unsystematic nature, a deficiency that UMMPIREE seeks to correct. Apart from not having common use of terms and definitions for assessment concepts and ideas, current assessments also tend to be conducted in accordance with procedures and practices that are unsystematic and unstandardized. Many of these unsystematic practices are found within the assessment planning stage.

4.2.1.1 Problem Statement: Unsystematic Assessment Planning

Assessment planning covers the procedures and practices involved in the setup of assessments. This ranges from determining the purpose for the assessment, to identifying the material and participants required, to designing the procedure for conducting the assessment. Not all assessments have clearly defined goals or hypotheses that can inform level of experimental control and research design. Assessments expediently planned under time and budgetary constraints may not use the appropriate tasks or measures for the constructs of interest. Assessments may not have been planned with the correct type or amount of resources and require study procedures to be modified midway through the assessments, affecting standardization. Many of these problems are due to assessment planners having different levels of knowledge and experience, and can result in assessments that do not yield useful conclusions and recommendations.

4.2.1.2 UMMPIREE Solution: An Assessment Planning Framework

To address such problems, an APF outlining the best practices in assessment planning was developed. It was intended to serve as a useful tool for both novice and expert assessment teams, providing specific guidance for the various stages of the assessment process. The APF was to be broadly applicable across the evaluation of technologies, studies of training interventions, introduction of new work procedures and policies, adoption of new organizational structures, and so forth. The process incorporated in the APF is consistent with best practices set forth by the Department of Defense (Bjorkman 2008; DA 2012) and NATO (DOD 2002).

Assessment is best thought of as a cyclic process within a Plan-Execute-Report framework. Reporting and critical analysis of outcomes feeds back into future planning efforts so that assessments can be improved on an iterative basis. Figure 4 unpacks the cycle as specified in the APF in more detail. The APF describes how assessment is accomplished across multiple phases: planning, execution (which includes data capturing, training aids, and monitoring), postprocessing data analytics, and cyclical improvement based on the observed results (see Appendix D). Instead of trying to develop a prescriptive one-size-fits-all approach, the APF presents the assessment team with a series of questions or issues for consideration at each stage of the process. In doing so, the APF specifies the decision points in the planning process that most critically impact the quality of the assessment. Responses to these questions will guide the assessment team toward the approach that best balances project-specific needs with opportunities for reuse.

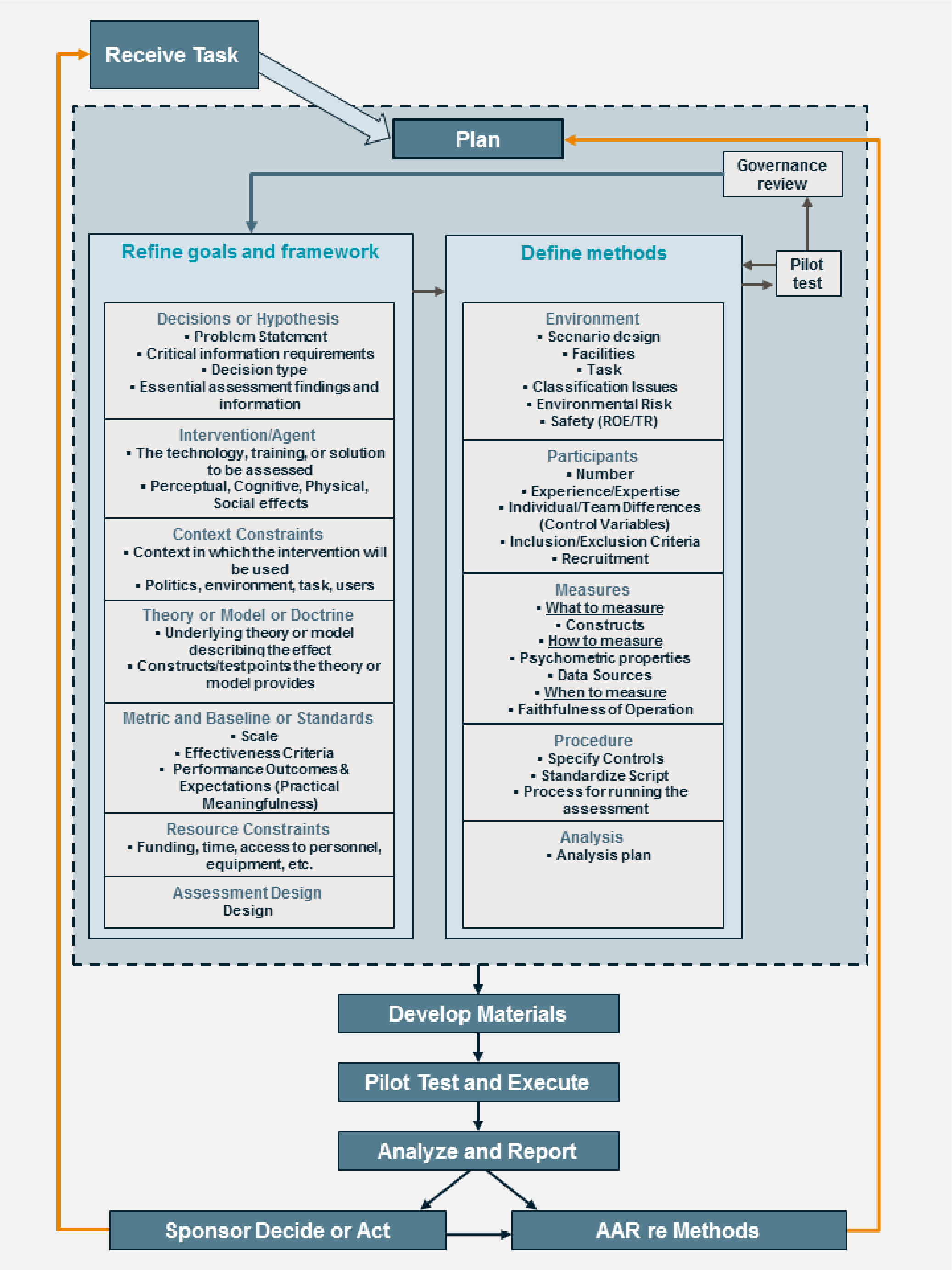

Receive Task
Plan
Governance review
Pilot test
Refine goals and framework
Decisions or Hypothesis
▪ Problem Statement
▪ Critical information requirements
▪ Decision type
▪ Essential assessment findings and information
Intervention/Agent
▪ The technology, training, or solution to be assessed
▪ Perceptual, Cognitive, Physical, Social effects
Context Constraints
▪ Context in which the intervention will be used
▪ Politics, environment, task, users
Theory or Model or Doctrine
▪ Underlying theory or model describing the effect
▪ Constructs/test points the theory or model provides
Metric and Baseline or Standards
▪ Scale
▪ Effectiveness Criteria
▪ Performance Outcomes & Expectations (Practical Meaningfulness)
Resource Constraints
▪ Funding, time, access to personnel, equipment, etc.
Assessment Design
Design
Define methods
Environment
▪ Scenario design
▪ Facilities
▪ Task
▪ Classification Issues
▪ Environmental Risk
▪ Safety (ROE/TR)
Participants
▪ Number
▪ Experience/Expertise
▪ Individual/Team Differences (Control Variables)
▪ Inclusion/Exclusion Criteria
▪ Recruitment
Measures
▪ What to measure
▪ Constructs
▪ How to measure
▪ Psychometric properties
▪ Data Sources
▪ When to measure
▪ Faithfulness of Operation
Procedure
▪ Specify Controls
▪ Standardize Script
▪ Process for running the assessment
Analysis
▪ Analysis plan
Develop Materials
Pilot Test and Execute
Analyze and Report
Sponsor Decide or Act
AAR re Methods

Fig. 4 Assessment process framework

In general, an assessment typically begins with the receipt of a specific task requirement, usually from the project sponsor or governing body. The assessment team then proceeds to plan for the assessment. Planning mainly involves refining the goals and overall framework for the assessment, which in turn help define the methods (further explained in the following sections). This initial plan is then pilot tested to determine task parameters and check manipulations and is iteratively refined to obtain a formalized plan. The formalized plans are then provided to the sponsor or governing bodies for review and sign-off. Upon sign-off, the assessment team would proceed to develop the requisite materials, enabling further pilot testing and eventually actual execution of the assessment and data collection. The data from the assessment are analyzed according to the plan, and the results reported. The outcome could be a review of the methods leading to modifications in the assessment plan or a decision or action by the sponsor that could result in further assessments. Next, the key stages in the assessment process are described in more detail.

Refine Goals and Framework

At the outset of planning, the assessment team has to clarify the goals and purpose of the assessment, which necessitates an understanding of the context and background of why the assessment was required. The following outlines the primary considerations when refining goals and the conceptual framework that guides assessment

Decisions or Hypothesis

The assessment's specific purpose will guide how it is planned and what the final deliverable will look like. For instance, assessments may be formative, summative, or conducted to establish a benchmark (Sanders 1994). Questions to help determine the purpose of the assessment include the following: *What is the problem statement? What are the decision maker's critical information requirements? What type of decision must be made? How will the assessment findings be used?*

Intervention/Agent

Assessments are conducted to understand some aspect of the environment, the individual, or the interaction of both of these (Matthews 2016). Identifying the subject matter or target of the assessment in the early stages of planning will help scope the assessment and allow estimation of the resources required for the assessment. Questions to be addressed include the following: *What technology, training, or intervention is to be assessed? What specific human performance outcomes (e.g., perceptual, cognitive, physical, or social) are expected to result from this intervention?*

Context Constraints

As with any initiative conducted within an organization, the assessment team needs an appreciation of organizational factors that can impact the assessment. Questions to help the assessment team anticipate how their work fits into the larger organization include the following: *What is the larger context in which the* *assessment will occur? What are the external constraints, including politics, the larger military environment, the tasks, and the intended users on the assessment process?*

Theory, Model, or Doctrine

Theories, models, and doctrine often provide indications of the constructs to include in assessments, the type of relationships expected among them, and the effects that may be observed. All of these help the assessment team put in place the task parameters and measures required to detect the anticipated effects. Questions that the assessment team should consider include the following: *What constructs should we measure? When should we expect to observe an effect? Is the proposed human augmentation a main effect, a mediated effect, or a moderated effect?*

Metric and Baseline or Standards

For assessment results to be useful and actionable, the assessment team should attempt to relate the assessments to existing baseline or standards. This entails using metrics that are linked to meaningful criteria. Questions that should be addressed include the following: *On what scale should human performance be measured? What are the criteria for effectiveness or proficiency? What human performance outcomes are considered practically meaningful?*

Resource Constraints

Resources such as funding, time, and participant availability, are required to conduct assessments and are not unlimited. Understanding these constraints would help the assessment team prioritize their requirements to maximize the resources they are able to obtain. Assessment teams would need to address questions such as the following: *How can we work within the likely constraints to design the best-possible assessment? How will these constraints impact our findings?*

Assessment Design

The assessment or study design relates to the research questions and hypotheses that the assessment can address and directly impacts the amount and type of resources needed. For instance, a within-subjects design enables good comparison between the control and experimental conditions, but requires participants to be repeatedly assessed, taking a longer period of time. Questions relevant to this component include the following: *What study designs can be adopted for the assessment with the amount of resources available? How does the study design affect the conclusions that can be drawn from the assessment results and decisions that the assessment can inform?*

Define Methods

After refining the goals and broad conceptual framework for the assessment, the team will need to specify the methods to be used in the assessment. These are the components that are often detailed in the methods section of a scientific research report. Each component raises questions that must be explicitly answered and justified to support the methodology chosen.

Environment

One of the most critical components of the assessment is the choice of stimuli and task scenario that are used to induce the behaviors or responses that represent the constructs of interest. These can range from scenarios in the real world to tasks simulated on experimental testbeds. Relevant questions to be considered when deciding on the environment include the following: *What environment, scenarios, and tasks would induce the effects we wish to study? What facilities, location, equipment, and technical support do we need?*

Participants

The ability of the assessment team to generalize assessment findings and study inferences depend on the number and type of participants. Although actual Soldiers are preferred as participants, they are not readily accessible. Questions to be considered include the following: *Who are the assessment participants (i.e., participant type and number)? How representative is the participant pool of the target population? Are the tasks and scenarios realistic for use by the participant pool?*

Measures

The measures used must be indicative of the constructs of interest as inappropriate measures threaten the validity of the study. Psychometric criteria for measure selection (O’Donnell and Eggemeier 1986; AERA et al. 2014) should be considered as well as questions like the following: *What to measure? How to measure? When to measure?*

Procedure

A good study procedure maximizes validity, reduces the number of confounds in the study, and can help preserve data quality. Questions to be addressed include the following: *What should the study procedure be (i.e., the data collection script, order of tasks)? What should the procedure for handling data be (i.e., storing and aggregating data)?*

Analysis

An analysis plan specified upfront before data collection allows the assessment team to identify any problems in the assessment that would prevent the required analysis to be conducted (e.g., having unsuitable comparison groups). Assessments can be adjusted at this planning stage to avoid making unnecessary compromises later. Questions that should be considered include the following: *What analyses are needed to answer the research questions? What types of data are needed for those analyses?*

4.2.2 Support for Multiple Stakeholders

In addition to outlining a planning framework to help assessments to be more systematic and standardized, UMMPIREE found that a major cause for unsystematic and unstandardized assessment lay in the fact that assessments are carried out by teams and personnel with varying level of expertise in assessments and who have different purposes and agendas for the assessments.

4.2.2.1 Problem Statement: Lack of Support for Assessment Stakeholders

When the practices and procedures in assessments have not been systematically thought through, crucial steps or information may be overlooked or under-considered. This can lead to assessments that are limited in the questions they address, and the assessments would not have adequately served various stakeholder needs. In addition, since assessment planning is typically conducted by the researchers and engineers, the assessments tend to be skewed toward addressing their needs.

4.2.2.2 UMMPIREE Solution: A Conceptual Architecture

An overall schema or architecture that outlines assessments from the perspectives of various stakeholders would help all involved to understand the multiple purposes and uses of the assessments. First, UMMPIREE initiated a description of multiple assessment stakeholders and their agendas that resulted from their respective roles (see stakeholder tiers in Appendix E). Next, the MAA was developed first as a conceptual architecture that depicted the assessment process from the perspectives of different assessment stakeholders. Although future development work will include building tools and apps and incorporating the assessment needs of other stakeholders, this initial effort focused on the “Assessment Planner” and “Event/Experiment Observer” (both in Tier 3 of Appendix E). The assessment needs and agendas of these stakeholders or users are described with the following examples.

Assessment planner:

- Human factors engineer with assessment expertise. Interested in what the innovative technology impacts and wants to know additional constructs that should be included in the assessment.
- Engineer who developed the innovative technology but has limited assessment expertise. Interested in what equipment is needed to set up assessments.

Event/experiment observer:

- Army lieutenant colonel (LTC) who is familiar with the technology. Interested in the initial assessment results to determine whether the innovation is worth further consideration.

In an example of an assessment of sniper robots that provide “smart” cover for squads, a human factors engineer may be interested to see if the new sniper robots improve situational awareness of Soldiers, while the engineer who developed the sniper robots may be more interested in assessing the functionality of the sniper robots under different conditions. In contrast, an LTC may wish to know if the sniper robots can provide “smart” cover for small squads and enable him to deploy fewer Soldiers for a mission. Depending on their use of assessments, different stakeholders would focus on different stages in the assessment process. Researchers and engineers may have a larger role in assessment planning and execution, while the LTC may take more interest in the assessment report of the results.

The MAA comprises 3 main system components with 2 underlying subsystems (Fig. 5). The system components follow the Plan–Execute–Report model, supporting compatibility with the APF. By expressing the assessment process in terms of system components, the MAA can serve as a framework for future tool or application development. Furthermore, the MAA presented assessments from the viewpoints of 3 stakeholders (Table 2). The MAA system/subsystem components and level of stakeholder interest in each component are presented in brief in the following (see Appendix F for details).

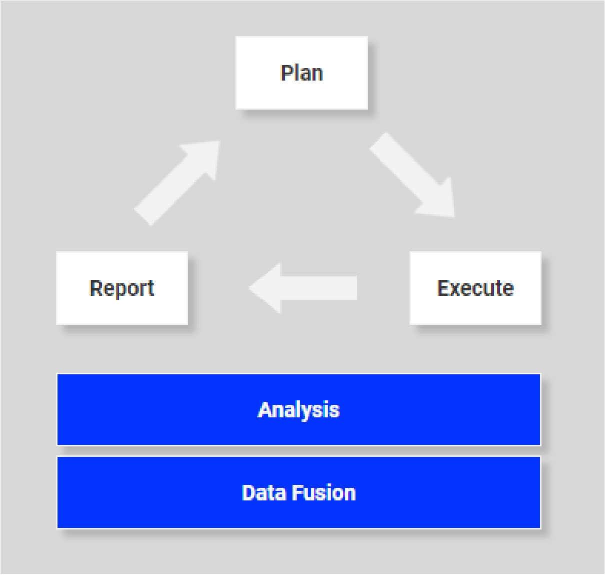


Fig. 5 MAA systems and subsystems

Table 2 Assessment needs of different users as represented in the MAA

MAA system/subsystem	User 1: human factors engineer	User 2: Army LTC	User 3: engineer
Plan Has a "Toolbox" module that contains resources for planning assessments (e.g., facilities, participants, staff, instruments, measures, and assessments). Has a "Smart Guide" module that can be developed into a "wizard". Has an "Assessment Plan" module that can be developed to generate an assessment plan for execution based on assessment purposes, requirements, needs, etc., specified in the "wizard".	Very high involvement	Very low involvement	High involvement
Execute Has a "Build" module that analyses and tests the Assessment plan generated from the "Plan" system. Has a "Verify" module that checks for conflicts in schedule and logistics. Has an "Execute module" that specifies the roles required to implement the assessment plan.	Very high involvement	Low involvement	Very high involvement
Report Has an "AAR" module that can be used for assessment debriefings. Has a "Summary Report" module that generates reports of the assessment.	Moderate involvement	Very high involvement	Moderate involvement

Table 2 Assessment needs of different users as represented in the MAA (continued)

MAA system/subsystem	User 1: human factors engineer	User 2: Army LTC	User 3: engineer
Analysis Supports data analyses and generation of assessment reports.	Very high involvement	Low involvement	Moderate involvement
Data fusion Supports the integration of multiple data and information assets.	Low involvement	Very low involvement	Moderate involvement

4.3 Activity 3

UMMPIREE's project objectives are to improve assessments and address the growing assessment needs of the military in view of emerging technologies and methods. While efforts from the previous activities provided the basic underpinnings to address project objectives, UMMPIREE is also seeking to develop solutions so that an organizational capability for systematic assessment development and validation can be built. These may include 1) tools or applications to assist assessment teams in planning and executing assessments that are systematic and standardized, and 2) methods for linking research that lacks the guidance from strong theories to its grounding in applications and field experience. To move forward with solution development, the UMMPIREE team focused on a specific domain to ensure that subsequent work addressed real-world challenges and needs.

Defining a Military Domain

Many of the challenges with assessments identified in the project relate to the fact that assessment teams in applied research settings typically comprise personnel with different levels of expertise in assessment and measurement, all of whom may have various agendas and interests. In addition, assessments are conducted within organizations and subjected to organizational pressures that often require assessment teams to compromise on certain areas of their assessment planning and execution. To improve assessments in such a context, the project team needs to develop tools that serve a variety of users, are practical, and support a level of interoperability.

4.3.1 Problem Statement: Need for a Domain That Exemplifies Assessment Challenges Identified

R&D occurs at different rates across military domains, and each faces a different set of challenges. The particular problems highlighted in this project are most apparent in domains that do not merely assess technological innovations and interventions per se, but also assess their effects on the human, including how humans interact with these. Such "human-centric" assessment is contrasted with the more "system-centric" assessment conducted in T&E.

A domain where human-centric assessments are regularly conducted was needed to focus the project team's efforts. With a domain for which to target future activities, the project team would be able to learn from the actual examples of how assessments have been conducted, and how specific assessment challenges have been dealt with. Having a domain would ensure that project efforts are grounded and provide a target community for any prototype of the tools developed. By engaging stakeholders within a domain, opportunities for tool implementation in that domain would increase.

4.3.2 UMMPIREE Solution: Focus on HMT

DOD confronts extraordinarily complex problems ranging from crisis relief to war. To succeed, it organizes its personnel and technologies into capabilities for distributed command and control, distributed mission operations, social informatics, crowd-sourced predictions, and participatory gaming. These hybrid (humans and machines) teams and organizations are designed to increase problem solving efficiency and effectiveness by improving situational awareness, sense-making, collaboration, and coordination over teams composed of only humans or only machines. ("Machines" includes robots, intelligent software agents, and other autonomous systems). Typically, successful human–machine collectives exploit the strengths of deliberative processing and instant recall characteristic of machines, or they leverage the generalization, learning, and adaptation proficiencies of humans.

The HMT domain is a relatively new domain that in only a few recent years has rapidly gained considerable research interest and the accompanying funding support from various agencies. Much of the research in the HMT domain resulted from the introduction of new technology in the form of robots. Unmanned aerial and ground vehicles and various other systems have not only revolutionized the way military operations are conducted, but also raised questions and concerns about the wider effects of these innovations. Human–machine partnerships are becoming increasingly ubiquitous, carrying out tasks that involve complex cognitive interactions and coupled decision making. However, there is currently no satisfactory science of human–machine teaming. Many of the measures and theories used and cited in this domain's research have been developed for traditional all-human teams, and there is increasingly a concern over the degree to which these measures and theories (e.g., measures of team effectiveness and cohesion, and team mental models) apply to the HMT domain.

Assessments in the HMT domain

Many of the assessments in the HMT domain involve technological innovations and augmenting interventions developed to enhance various aspects of human performance, ranging from sensing, to information processing, and physical abilities. Table 3 provides an illustrative list of the technologies that have been evaluated in the HMT domain.

Table 3 Examples of training and augmentation interventions

Category	Example
Information systems/aids	Google translate Commander's virtual staff IBM Watson oncology advisor Blue Force Tracker Pilot's associate
Sensors	SmartGlasses (augmented reality) Smart helmet (daqri.com); includes 4-D work instructions, thermal vision, data visualization, remote expert
Chemical/genetic	Erythropoietin and other performance enhancing drugs and associated genetic mutation/variability Amphetamines
Physical (augmentation)	Exoskeleton Transcranial IR laser stimulation
Weapons	Tasers Sound
Other	User experience improvements over existing user interfaces

The HMT domain also provides case studies that exemplify the challenges toward which UMMPIREE is directed. For example, trust is a key construct for HMT, but assessment methods are limited and in many cases borrowed from other domains such as human–human trust and human–automation trust (Schaefer 2013). It remains to be determined whether key constructs from these domains are valid for HMT. The dynamic, unstructured nature of many human–robot interaction scenarios (Desai et al. 2012) and anthropomorphic attributions of autonomous systems (Waytz et al. 2014) may limit construct generalization. A scale that predicts optimization of trust in human–human interaction must be revalidated for the HMT domain. Understanding HMT may also require assessment of novel constructs. Popular culture may produce attitudes and emotions unique to the domain, operationalized for example in scales for negative attitudes toward robots and robot anxiety (Nomura et al. 2008). Trust may also have implicit, unconscious elements requiring behavioral assessments. Another novel construct is the human attribute of theory of mind (ToM), which enables humans to cooperate and team with other humans (Scassellati 2001; Hiatt et al. 2011; Streater et al. 2012; Wiltshire et al. 2013). ToM refers to the human ability to infer the mental states (e.g., thoughts, intent) of another human from their behaviors (e.g., speech, facial expression, and gestures). Assessments are required to determine the degree to which robots have to convey the "presence" of ToM to foster trust and teaming with the human.

Apart from novel constructs, HMT trust research also encompasses the study of new phenomena that require measures that go beyond what is currently available. The Uncanny Valley phenomenon (Mori 1970) was discovered as a result of human–robot interaction research and prior to advances in technology and robotics was unknown. In a rescue scenario, anthropomorphic robots tending to trapped victims were viewed as "creepy" and not comforting (Murphy et al. 2004), suggesting that human–robot interaction and trust can be affected by the extent to which degree of robot similarity to human appearance approaches the Uncanny Valley (Minato et al. 2004). Existing assessment methods and measures are not suitable for the study of such new phenomena.

The following trust assessments for HMT also illustrate the specific challenges listed in Section 2.3.:

1) Constructs relevant to trust are numerous and diverse; for example, a meta-analysis (Hancock et al. 2011) identified multiple constructs within each of the categories of robot performance, attributions, operator ability, personality, and environmental features.

2) Qualitatively different assessments are available, including behavioral, physiological, and self-report measures (Waytz et al. 2014).

3) Assessments must cover humans, machines, and their dynamic interaction. This should include the effects of the technology or machine of interest as well as the combined effects of that technology and other existing systems that the user works with.

4) As Table 3 indicates, there are multiple contexts and purposes for assessing trust that will refer to different timespans. Contrast, for example, trust as an element of immediate user experience versus trust as a long-term predictor of performance gains from training or augmentation.

5) There will also be multiple stakeholders with different perspectives, including end users, commanders, mission planners and analysts, as well as researchers. The various systems listed in Table 3 will also require assessment in different locations. For example, sensors may be tested both in controlled laboratory settings and in field operations.

4.4 Activity 4

The tools and products that will be developed from UMMPIREE will help promote systematic and standardized assessments. They are targeted for use by assessment teams in domains, such as the HMT domain, that face many of the assessment challenges identified in the project. For the tools to be viable and functional, UMMPIREE needs to incorporate instances and details of the assessment experiences of HMT assessment teams into the baseline. This activity used inputs from experts in the domain to develop tools that support integration of findings from multiple, potentially disparate, subdomains of research, and for visualizing metadata for subdomains.

SME Input on Baselines

To augment the baselining of current assessment practices and challenges with real-world examples, UMMPIREE engaged SMEs in HMT assessment to solicit their inputs.

4.4.1 Problem Statement: Need SME Contribution and Buy-in to Project and Solutions

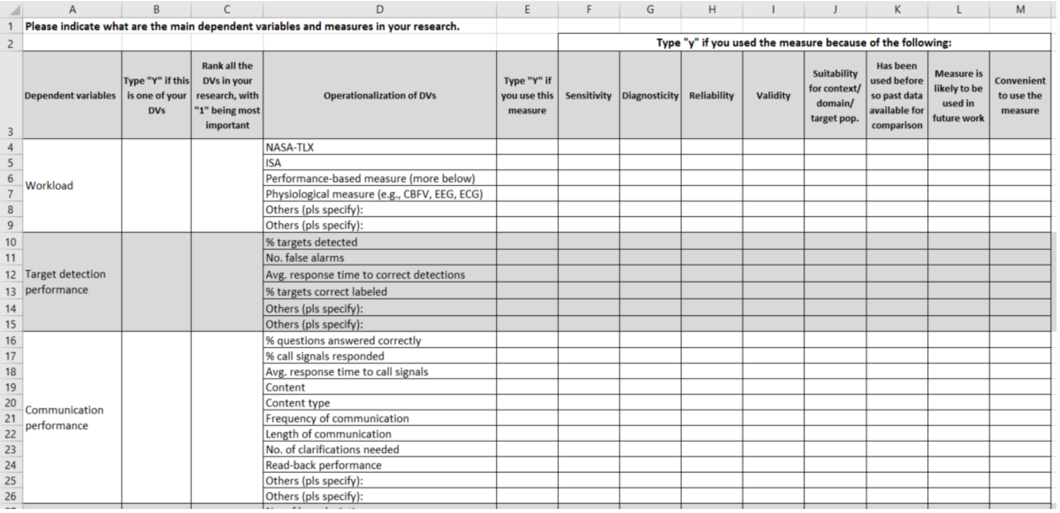
SMEs in HMT assessment are found in various research labs across the US military services (e.g., ARL, Air Force Research Laboratory, and Naval Research Laboratory). They routinely conduct assessments to evaluate the effects of augmenting technology and interventions on operators and teams and thus potential users of the solutions developed in UMMPIREE. SME involvement will lend credibility to many of the larger objectives of the UMMPIREE project. However, they are a heterogeneous group, operating in labs with different capabilities and resources and pursuing different immediate goals. SME availability for involvement and contribution to the UMMPIREE effort and likelihood of adopting UMMPIREE’s solutions may require reaching out to engage these individuals and demonstrating the utility of UMMPIREE to multiple programs of research.

4.4.2 UMMPIREE Solution 1: SME Interviews

In addition to acquiring richer content for the baseline, the UMMPIREE team conducted face-to-face semi-structured interviews with SMEs to obtain their support for the project. The project team wanted to engage SMEs from the various services and managed to interview several Army and Air Force researchers. Face-to-face knowledge elicitation was utilized to maximize SME engagement and display commitment to the overarching objectives of the UMMPIREE project. Once the introductions between SMEs and UMMPIREE team members were completed, the UMMPIREE team members gave a short overview detailing the objectives of the interview.

SMEs were interviewed on the type of research they were currently pursuing or completing, items that they may need to consider as their work progressed, and the research challenges that they currently face or anticipate facing in the future. Some of the specific questions focused on items such as the following: 1) areas of interest, 2) units of analysis, 3) independent and dependent variables, measures, and factors, 4) typical funding sources, 5) constructs, theories, and models used or investigated, and 6) any local or global barriers that impede the progress of their research (see Appendix G for list of interview questions). The answers to these questions were manually recorded and juxtaposed with other SME responses to develop common thematic notions.

Additionally, a postinterview questionnaire was circulated to give the SMEs a chance to supplement their answers and rank some of the factors discussed in the face-to-face interview (Fig. 6). After the interview, the UMMPIREE team collated manual notes and SME questionnaires in a consistent and shareable format. The objective of this effort was to acquire valuable SME input and make this input machine-readable to facilitate the development of an SME database tool or future tools as necessary.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Please indicate what are the main dependent variables and measures in your research.**												
2						**Type "y" if you used the measure because of the following:**							
3	**Dependent variables**	**Type "Y" if this is one of your DVs**	**Rank all the DVs in your research, with "1" being most important**	**Operationalization of DVs**	**Type "Y" if you use this measure**	**Sensitivity**	**Diagnosticity**	**Reliability**	**Validity**	**Suitability for context/ domain/ target pop.**	**Has been used before so past data available for comparison**	**Measure is likely to be used in future work**	**Convenient to use the measure**
4	Workload			NASA-TLX									
5				ISA									
6				Performance-based measure (more below)									
7				Physiological measure (e.g., CBFV, EEG, ECG)									
8				Others (pls specify):									
9				Others (pls specify):									
10	Target detection performance			% targets detected									
11				No. false alarms									
12				Avg. response time to correct detections									
13				% targets correct labeled									
14				Others (pls specify):									
15				Others (pls specify):									
16	Communication performance			% questions answered correctly									
17				% call signals responded									
18				Avg. response time to call signals									
19				Content									
20				Content type									
21				Frequency of communication									
22				Length of communication									
23				No. of clarifications needed									
24				Read-back performance									
25				Others (pls specify):									
26				Others (pls specify):									

Fig. 6 Snapshot of the initial version of the postinterview SME questionnaire

SME Interview Results

Common categories and themes emerged from the SME interviews and associated post-interview questionnaires. Unsurprisingly, there were numerous recurring themes that emerged as the qualitative SME data were parsed. Specific examples of the most salient challenges and issues mentioned by the preliminary SME interviews are as follows.

- Access to Actual Operators. All SMEs mentioned that they have very limited access to actual operators. This conundrum has resulted in the need to simplify tasks for the participant pool to which they have access (i.e., university students). Thus, questions regarding the validity of findings are warranted and prevalent. To some SMEs, this lack of access reflected the services' perspective that the research labs are a resource providing research solutions. Hence, researchers should not "expend" resources and request support such as access to actual operators as study participants.

- Modifying Tasks for Novice Participants. Adding to the earlier point, the SMEs frequently have to modify tasks for novice participants. This potentially threatens external validity because some measures have to be invented (e.g., subtract points if did not see the red truck, or workload measured by performance on artificial secondary task like performing math while detecting targets). The task design is also currently constrained by testbed/environment.

- Task-Specific Findings. Many research findings can be task-specific (e.g., manipulations/stimuli are task-specific). This makes it difficult to generalize the findings across multiple domains. The tasks may also not really reflect actual operations (simplified for the novice sample). There is often little to no funding available to replicate preliminary results, making it difficult to check the robustness of findings.

- Construct Validity and Terminology Implications. With the development of potentially new constructs, issues often arise. For example, is human–machine trust the same as human–human trust? Team cognition is also another example where construct validity issues persist. It is possible that prevalent terminology has shaped assessment methods. For example, in using the term "team situational awareness", there may be a tendency to use the Situational Awareness Global Assessment Technique (Endsley 2000) as the measure. However, the validity of any measure should not be assumed without checking first.

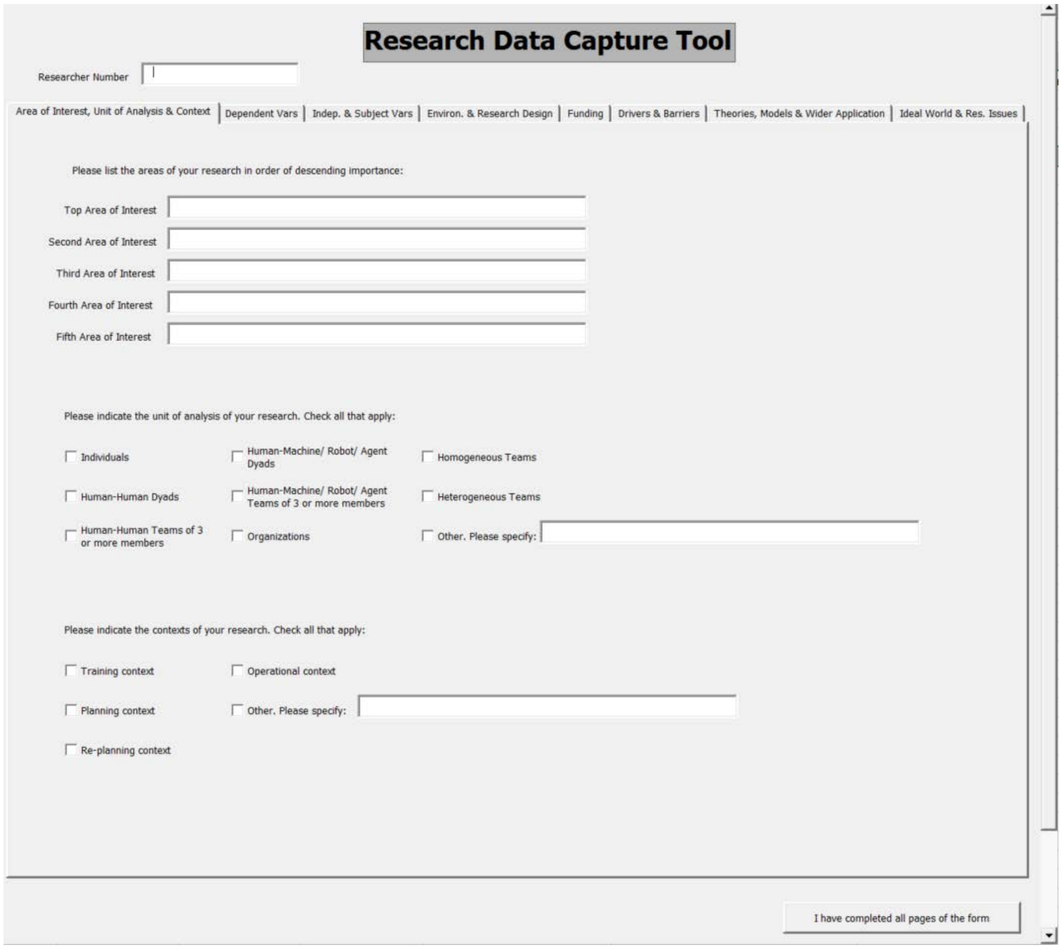
- Limited Realistic Testbed/Environment. Unfortunately, real testbeds for assessing teaming activity or realistic/high-fidelity testbeds are scarce and seldom representative enough to play out sequences of mission activity that reflect actual operations. However, the SMEs mentioned that they may not want their research to be constrained by existing systems; instead, they recognize that their research should contribute to decisions on future testbeds/environments. Nevertheless, when testbeds and tasks need to be created, researchers may not have access to SMEs who can help, reiterating the service's perspective of research labs that they should be the "resource that serves the service" instead of seeking to utilize the service's resources.

- Funding and Collaboration. Research labs are not directly given funds to attract research partners (e.g., universities) to help with some of these problems (e.g., lack of testbed/SMEs). Instead, they have to offer an exchange of services, and such arrangements do not always mutually benefit each party. Additionally, university research also tends to be driven by funding availability, which further exacerbates the problem.
- Discordance with Management. Senior management does not always appreciate the complexities of research (e.g., "Why can't we find the best [i.e., universally pervasive] workload measure?" or "What is so hard about being vigilant and monitoring a display for targets; the automation is already helping with this?"). It is possible that there is a disconnect between the expectations of senior management and researchers regarding what research can address.
- Changing Research Priorities. There was a common sentiment that management will change the research priorities before the research is completed. Management also tends to not appreciate the need to "dig deeper" to gain a thorough understanding of the research thrust (e.g., "Individual differences need to be investigated, but this requires a lot of time and money to do it properly" or "The current way of doing research is not conducive for this type of research question to be studied.") Lastly, political agendas also occasionally come into play and may introduce factors that thwart good research.

4.4.3 UMMPIREE Solution 2: Research Database Capture Tool

Responses from the researchers sampled during the baselining phase revealed several common themes in HMT research. With more data points, a more comprehensive picture of the state of assessments in HMT research can be built. This can uncover system-wide issues and challenges, broad funding trends and research direction within the domain. However, although the SME interviews yielded rich information and allowed the project team to better engage SMEs, they were resource-demanding. To supplement the SME interviews, the postinterview questionnaire was further developed into the Research Data Capture Tool (RDCT), which was more widely disseminated to obtain inputs from more SMEs.

The RDCT included a graphical user interface that collected SME input and was also designed to encourage the respondent to adopt a systematic approach to assessment planning and execution. This tool automated and facilitated data collection to enable data capture from more researchers and over a longer period of time. It captured SME data and stored it in text format (*.csv) in a relational database to enable easy linkage to other database models. These data, which comprised information on pertinent details of assessments such as constructs assessed, measures of these constructs, environments and tasks in the assessment, and the like, support recommendations in assessment planning decisions for future research in the same area, thereby promoting standardization of assessments (see example screenshot in Fig. 7)



Research Data Capture Tool

Researcher Number

Area of Interest, Unit of Analysis & Context
Dependent Vars
Indep. & Subject Vars
Environ. & Research Design
Funding
Drivers & Barriers
Theories, Models & Wider Application
Ideal World & Res. Issues

Please list the areas of your research in order of descending importance:

Top Area of Interest

Second Area of Interest

Third Area of Interest

Fourth Area of Interest

Fifth Area of Interest

Please indicate the unit of analysis of your research. Check all that apply:

□ Individuals
□ Human-Machine/ Robot/ Agent Dyads
□ Homogeneous Teams
□ Human-Human Dyads
□ Human-Machine/ Robot/ Agent Teams of 3 or more members
□ Heterogeneous Teams
□ Human-Human Teams of 3 or more members
□ Organizations
□ Other. Please specify:

Please indicate the contexts of your research. Check all that apply:

□ Training context
□ Operational context
□ Planning context
□ Other. Please specify:
□ Re-planning context

Fig. 7 RDCT screenshot

Further Use of SME Data

The content from the SME interviews and RDCT not only added to the baselining, but also served as a resource for future solutions. For example, a future Web-based tool to assist in assessment planning could draw recommendations for measures and tasks to use for constructs from this SME database containing records of what had been done in previous assessments with the same constructs. This would help encourage standardization of construct operationalization to an extent. Serving also as a repository for information on research practices and challenges, including sources of funding and the like, this SME database would also be a resource for future assessment planners.

4.4.4 UMMPIREE Solution 3: Research Visualization Tool (RVT)

The data from the baselining interviews revealed that the SMEs' research programs were very much related, as there were similar research questions investigated and common constructs and measures used. The RVT was developed to visualize the metadata of their research, showing the links among the constructs and research areas as well as links among the SMEs and to other common researchers. Figure 8 is an example of this visualization. Such visualizations can provide a quick visual analyses of research hotspots in the domain for researchers as well as program managers and funding agencies.

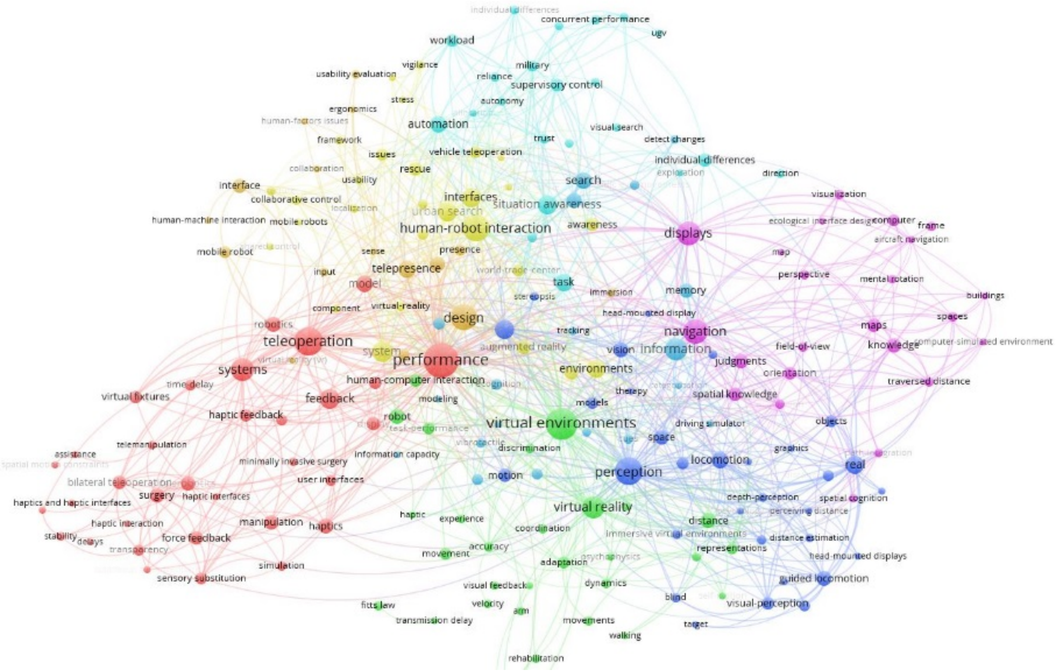


Fig. 8 Example of RVT

Providing Preliminary Data for the Linking Approach

The baseline data served as pilot data to test out the linking approach that would be developed in the next phase of the project. The data, which are multivariate and multilevel, allow the team to explore various techniques and criteria for linking to show meaningful relationships among unstandardized assessments from different labs that had assessed similar constructs using different on measures, tasks, and environments.

5. Phase II (Year 2): Linking Research

In many studies, different measures have been used for the same construct in various domains and environment, which has resulted in challenges in generalizing research findings. This source of confusion has impeded the understanding of constructs and their interrelationships. UMMPIREE's effort to address this issue involves developing a mathematical method to link fragmented research in addition to facilitating systematic assessments. Toward this end, the team would first require a firm understanding of how various assessment components (e.g., construct, measure, domain, and environment) relate to and influence each other. A conceptual basis or rationale for their linking to justify applying a mathematical method to link these would be required. Subsequently, the technical infrastructure would need to be developed to house information on assessment components gleaned from thorough literature reviews of research in the domain.

Anticipated activities within the linking phase are as follows:

1) Derive a conceptual basis for linking.
2) Formulate a mathematical approach for linking.
3) Identify and develop prototypes of products/apps with the supporting technical infrastructure for housing data elements.

From these activities, UMMPIREE seeks to be able to derive relationships among research studies conducted within the HMT domain that will be useful in future assessments. For instance, the linking can identify the most-frequently used and appropriate measures for a certain construct assessed in a particular environment/task.

5.1 Conceptual Basis for Linking Approach

One of the critical elements of our overall project concerns the issue of theory development and theory integration. It is clear that there are useful individual theoretical constructs that deal with some portion of individual elements of the overall project. However, what we do not yet possess is an articulated descriptive theory of the link between domains, constructs (affect), and measures. Neither do we possess a full understanding of the causal linkages between these disparate descriptive elements. The purpose of the work under the theory element has been to begin to satisfy this requirement, to identify issues and shortfalls associated with the present state of the art, and to identify fruitful avenues of future progress. In association with the network approach, a fully detailed architecture, so envisaged, would not only help with the present UMMPIREE project, but could well be on the way to systematic progress in all of the behavioral sciences.

First Specification

The first effort involved in such an ambitious but important project element has been the development of an overall descriptive framework that links work domains to functional states of the operator such as workload, stress, and situation awareness, and hence to outcome performance (Hancock et al. 2017). This component has been accomplished and is epitomized in the following tripartite illustration (Fig. 9), using "affect" (construct) as a shorthand for functional state. A fuller exposition of this work can be found in Hancock et al. (2017). Briefly, we have linked our own extension of the Perrow domain taxonomy (Perrow 2011) to affective states captured within the 3-D space of processing form (implicit vs. explicit), hedonic attribution (aversive vs. attractive), and memory differentiation (prospective vs. retrospective). We have then drawn linkages to verifiable measures of outcome response capacity (Fig. 9). This provides a framework for our future systematic evaluations (Hancock et al. 2017).

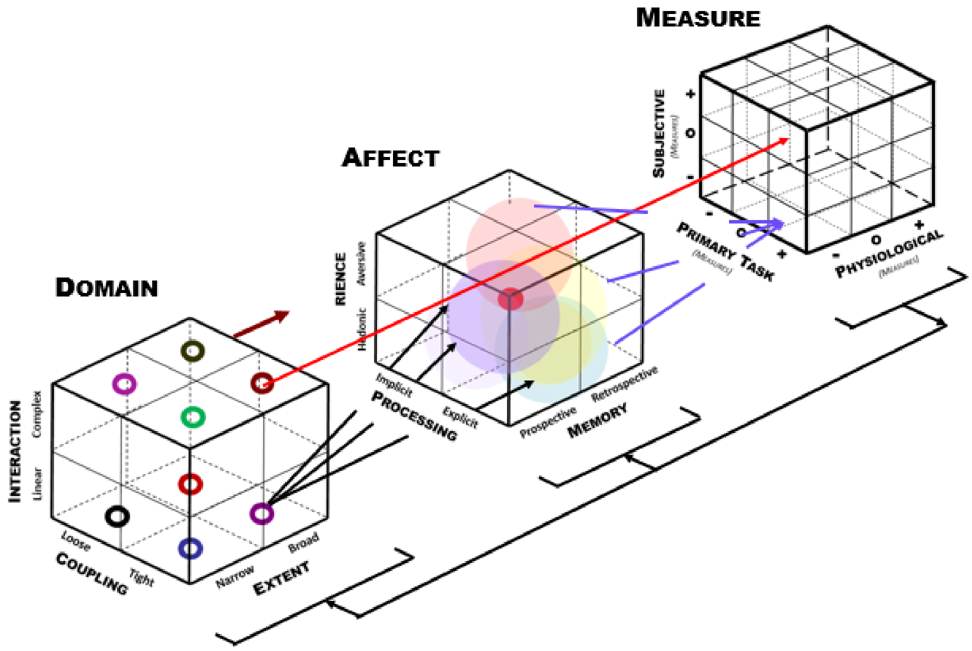


Fig. 9 Three-part differentiation of components of domain, affect (construct), and measure. Mediations and moderations are identified for future exploration.

Future Plans

The following plans are more fully explicate of the dimensions of domains, seeking to specify this crucial explication of the contextual environment of performance, now in much greater detail. The team also intends explore the ways in which current models and theories of human performance can be integrated into a much wider and broad-spread architecture, even if such an architecture is descriptive in its earliest forms. Next we will look to engage in a similar, in-depth focus on the specification and linkages of the identified affective states. The final step in the next phase of development is greater specification of the performance measurement taxonomy. The latter will complete the full descriptive picture and permit the evaluation of moderating and mediating influences in preparation for a wider and causal representation. We anticipate that the latter can help guide R&D by identifying shortfalls and gaps to be addressed by supported experimental, theoretical, and synthetic exploration.

5.2 Mathematical Basis for Linking Approach

While the conceptual basis for linking postulates mapping at the conceptual level, the mathematical approach is but one example of how that can be instantiated. This mathematical approach seeks to identify connectivity and similarity across constructs and among researchers.

Set Theory

Associated with each construct within a domain namespace is a set of measurable observables. Applying set theory to the sets of observables seems a natural way to discuss a mathematical representation of similarity across constructs. In comparing observable sets, the team will avoid any discussion of the role of observables/measures, either as describing or defining constructs. Instead, we accept the association as a given and reason from there. In comparing observable sets, we have investigated both relative and absolute comparisons. For relative comparison, we are using the Jaccard distance between 2 sets, but we recognize that this does not provide sufficient information. For example, we can imagine sets A and B, where $A \subseteq B$, and the Jaccard distance is any number between 0 and 1. To augment the relative comparison, we look to the relative complement to provide an absolute comparison. Combining both Jaccard distance and relative complement gives us a way to relate sets of observables and, therefore, constructs.

Associating constructs with sets of observables describes a partial ordering on the constructs. We will investigate if these threads of constructs can provide insight into the taxonomy described.

Network Theory

We may regard both the results of the RVT and the set theoretic approach to constructs and observables/measures as creating social networks that can now be analyzed using current (and future) social network analysis (SNA). There will be no direct connections between researchers or between constructs. Instead, one is related to the other through key words (researchers) or observables (constructs). We have just begun this work and are particularly interested in how SNA may identify influencers among constructs as well as possibly link concepts across researchers. Later work may include applying both these areas to individual data collected.

5.3 Technical Infrastructure for Housing Data Elements

As the raw data elements are identified, we will be able to develop an information model of how the data is structured, related to other data, and how the data can be combined to link to constructs. That information model will be the basis for the technical design decisions of the infrastructure to manage the data. The technical design will account for storing, querying, analyzing, visualizing, and manipulating the data. Database systems come in many variants (object, relational, graph, key-value, triplestores, etc.) It is important to understand the data being stored, how they will be accessed and queried, and the use cases of working with the data before deciding on the type of database and the tools associated with managing the data.

Other technical design decisions also depend on solidifying the use cases, such as whether the data will be stored in the cloud, stored on mobile devices, include massive amounts of storage, have semantic implications, require dynamic or static links, and the like. Some examples include the following: 1) if the relationships among the data are expected to change, ontologies are more flexible than relational databases, and 2) if the data will be updated/accessed by many users simultaneously, then a database with transactions would be useful to avoid users modifying data during queries.

Some design goals for the technical infrastructure include the following:

- Be accessible from anywhere at any time.
- Be modular and easily changed as our theory is adjusted and new data sets are provided.
- Have an easily changed structure since we expect our information model will change over time.

- Include a mechanism to link data and have additional metadata on the links themselves.
- Store large amounts of data gathered from research efforts.
- Have good performance for queries, data insertion, and data modification.

Use cases and requirements for the technical infrastructure will be developed toward the later part of Phase II. This will allow the team to collect sufficient data, develop an information model, and work with SMEs and future tool users on their user stories.

6. Phase III (Year 3): Cross-Pollinating

Objectives of the Cross-Pollinating Phase include extending the tools developing in the Baselining Phase and the linking approach and infrastructure developed in the Linking Phase to another domain. This includes engaging stakeholders in other domains and providing the resources and information acquired from the earlier phases to help improve assessments in that domain. The effort may also be extended to other users of assessments within the HMT domain.

Proposed activities for Phase III are as follows:

1) After work with the HMT domain, identify new related research domain to which UMMPIREE's work can be extended readily.
2) Modify existing tools and products for the new research domain.
3) Extend project initiatives to other stakeholder tiers in the HMT domain.

7. Conclusion

Assessment plays a critical role in maintaining and improving human performance effectiveness in the US Army. Realizing the benefits of new technology requires valid assessments of augmentation and training interventions. However, the unique nature of the military context for assessment raises pressing challenges. The size and diversity of the Army limits communication and synergies between different research teams. Similarly, the needs of different stakeholders in the assessment process may not be recognized. Assessment is intrinsically complex because numerous, sometimes ill-defined constructs shape performance, and the definition and operationalization of constructs varies across different military contexts. The challenge is accentuated by the Army's role as an early adopter of cutting-edge technologies for performance enhancement, requiring new forms of assessment and accommodation of rapid technological change.

UMMPIRREE aims to enhance the institutional value of assessment as a platform for enhancing human capabilities organization-wide. The project seeks to identify specific issues and limitations with current assessment practices and develop conceptual frameworks as well as practical software tools for addressing those issues. The products of UMMPIREE are intended to enhance both the work of individual research teams and organizational resources for enhancing capability.

The project is based on a 3-phase approach. Phase I, which is nearing completion, is a baselining phase that seeks to build on a systematic understanding of the current state of the art in Army assessment and needs for improvement. It has defined HMT as a focal domain for developing systematic practices for enhancing the planning, execution, and communication of assessments, especially in the contexts of augmentation and training. Phase II will focus on strategies for keeping pace with evolving technologies and innovations in methodology. These include using mathematical approaches to link assessments from different studies that can support software aids, such as visualizations of related research thrusts. Phase III will cross-pollinate products of the research across additional application domains so that they so that they are readily adaptable to various assessment contexts in support of an organization-wide initiative.

This report has detailed the activities of Phase I, how they support the overall goals of UMMPIREE, and associated software products and tools. Activity 1 aimed to support systematization of conceptual frameworks for assessment by clarifying basic principles, specifying a common lexicon for assessment research, and developing a CAM for operationalizing constructs. Activity 2 elaborated from this foundation to develop a conceptual framework and architecture for guiding assessments in multiple contexts, with Web software tools for transitioning concepts to practical application. Activity 3 established the suitability of the HMT domain as a testbed for UMMPIREE efforts due to its present and future military significance, the centrality of fast-changing technology, and its instantiation of the various challenges to valid, generalizable assessment. Activity 4 established a baseline for existing assessment practices in HMT by soliciting inputs from domain SMEs. On this basis, tools were developed for capturing the knowledge base of experts and visualizing the state of research, as a stepping stone toward Phase II.

Taken together, the Phase I efforts provide an integrated solution to the current lack of standardization and lack of systematic assessment procedures by developing flexible conceptual models as well as practical software tools that can be applied organization-wide. Future phases of UMMPIREE will build on these accomplishments to maximize the benefits of the program for optimizing capability enhancement throughout the Army.

8. References

American Educational Research Association (AERA), American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing. Standard for educational and psychological testing. Washington (DC): AERA; 2014.

Bjorkman E. Joint test and evaluation methodology (JTEM) overview. Presented at a meeting of the NDIA Systems Engineering Division Developmental Test and Evaluation Committee; 2008 June 23 [accessed 2016 July 5]. http://www.ndia.org/Divisions/Divisions/SystemsEngineering /Documents/Committees/Developmental%20Test%20and%20Evaluation %20Committee/2008/June%20Committee%20Meeting/6-24-08%20NDIA %20DTE%20Committee%20Charts.pdf.

Boldovici JA, Bessemer DW, Bolton AE. The elements of training evaluation. Alexandria (VA): US Army Research Institute for the Behavioral and Social Sciences; 2002.

Charlton SG, O'Brien TG. Handbook of human factors testing and evaluation. 2nd ed. Mahwah (NJ): Lawrence Erlbaum; 2001.

Chen JY, Procci K, Boyce M, Wright J, Garcia A, Barnes M. Situation awareness-based agent transparency. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2014 Apr. Report No.: ARL-TR-6905.

[DA] Department of the Army. Systems approach to training: evaluation. Ft Monroe (VA): Training and Doctrine Command (TRADOC); 2004. TRADOC Pamphlet No.: 350-70-4.

[DA] Department of the Army. The US Army learning concept for 2015. Ft Monroe (VA): Training and Doctrine Command (TRADOC); 2011. TRADOC Pamphlet No.: 525-3-0.

[DA] Department of the Army. Army studies and analyses. Army regulation no. 5–5. Washington, DC: DA; 2012 [accessed 2016 July 5]. http://www.apd .army.mil/pdffiles/r5_5.pdf

[DA] Department of the Army. The US Army operating concept: win in a complex world. Ft Eustis (VA): Training and Doctrine Command (TRADOC); 2014a. TRADOC Pamphlet No.: 525-3-1.

[DA] Department of the Army. The US Army human dimension concept. Ft Eustis (VA): Training and Doctrine Command (TRADOC); 2014b. TRADOC Pamphlet No.: 525-3-7.

[DA] Department of the Army. The Army 2025 vision: strategic advantage in a complex world. Washington (DC): DA; 2015. http://usacac.army.mil /publication/the_army_vision.

[DOD] Department of Defense. NATO code of best practice for C2 assessment; 2002. ISBN 1-893723-09-7.

[DODD] Department of Defense Directive 7045.20. Capability portfolio management; 2008.

Desai M, Medvedev M, Vázquez M, McSheehy S, Gadea-Omelchenko S, Bruggerman C, Steinfeld A, Yanco H. Effects of changing reliability on trust of robot systems. Proceedings of the 7th ACM/IEEE International Conference Human-Robot Interaction. New York (NY): Association for Computing Machinery (ACM); 2012. p. 73–80.

Edward JR, Bagozzi RP. On the nature and direction of relationship constructs and measurement. Psychological Methods. 2000;5(2):155–174.

Ellman J, Samp L, Coll G. Assessing the third offset strategy: a report of the CSIS international security program. Washington (DC): Center for Strategic and International Studies (CSIS); 2017.

Endsley MR. Direct measurement of situation awareness: validity and use of SAGAT. Situation Awareness Analysis and Measurement. 2000;10.

Freeman J, Stacy W, Olivares O. Assessment for learning and development in virtual environments. In: Schmorrow D, Cohn J, Nicholson D, editors. The PSI handbook of virtual environments for training and education: developments for the military and beyond. Westport (CT): Praeger Security International (PSI); 2009. p. 236–250.

Funke GJ, Knott BA, Salas E, Pavlas D, Strang AJ. Conceptualization and measurement of team workload: a critical need. Human Factors. 2012;54(1):36–51.

Goodwin LD, Leech NL. The meaning of validity in the new. Measurement and Evaluation in Counseling and Development. 2003;36(3):181–191.

Hancock PA, Billings DR, Schaefer KE, Chen JY, De Visser EJ, Parasuraman R. A meta-analysis of factors affecting trust in human-robot interaction. Human Factors. 2011;53(5):517–527.

Hancock PA, Burford CW, Teo GW, Reinerman-Jones LE. The conceptual basis for a mapped domain-affect-performance topography; 2017. Chapter in progress.

Hawley JK. Training and testing: a complex and uneasy relationship. The ITEA Journal of Test and Evaluation. 2006 Dec/2007 Jan:34–40.

Heiman GW. Research methods in psychology. New York (NY): Houghton Mifflin Company; 2002.

Hiatt LM, Harrison AM, Trafton JG. Accommodating human variability in human-robot teams through theory of mind. Proceedings of the International Joint Conference on Artificial Intelligence. 2011;22(3):2066.

Krumm S, Hertel G. Knowledge, skills, abilities and other characteristics (KSAOs) for virtual teamwork. In: Derks D, Bakker AB, editors. The psychology of digital media and work. Hove (UK): Psychology Press; 2013. p. 80–99.

Matthews G. Multidimensional profiling of task stress states for human factors: a brief review. Human Factors. 2016;58(6):801–813.

Matthews G, Reinerman-Jones L. Workload assessment: how to diagnose workload issues and enhance performance. Santa Monica (CA): Human Factors and Ergonomics Society; in press.

Minato T, Shimada M, Ishiguro H, Itakura S. Development of an android robot for studying human-robot interaction. Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems: Innovations in Applied Artificial Intelligence; 2004. p. 424–434.

Mori M. The uncanny valley. Energy. 1970;7(4):33–35.

Murphy RR, Riddle D, Rasmussen E. Robot-assisted medical reach back: a survey of how medical personnel expect to interact with rescue robots. Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication; 2004; p. 301–306.

National Council on Measurement in Education glossary of important assessment and measurement terms [accessed July 2017]. http://www.ncme/NCME /Resource_Center/Glossary/NCME/Resource_Center/Glossary1.aspx?.

Nomura T, Kanda T, Suzuki T, Kato K. Prediction of human behavior in human-robot interaction using psychological scales for anxiety and negative attitudes toward robots. IEEE Transactions on Robotics. 2008;24(2):442–451.

Nunnally JC, Bernstein IH. Psychometric theory. New York (NY): McGraw-Hill Inc.; 1994.

O'Donnell RD, Eggemeier FT. Workload assessment methodology. In: Boff KR, Kaufman l, Thomas JP, editors. Handbook of perception and human performance. Vol. 2. Cognitive processes and performance. Oxford (UK): John Wiley; 1986. p. 1–49.

Perrow C. Normal accidents: living with high risk technologies. Princeton (NJ): Princeton University Press; 2011.

Plake BS, Wise LL. What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing? Educational Measurement: Issues and Practice. 2014;33(4):4–12.

Sanders, AF. Simulation as a tool in the measurement of human performance. Ergonomics. 1991;34(8):995–1025.

Sanders, JR. The program evaluation standards: how to assess evaluations of educational programs. Kalamazoo (MI): The Joint Committee on Standards for Educational Evaluation; 1994.

Scassellati BM. Foundations for a theory of mind for a humanoid robot. Cambridge (MA): Massachusetts Institute of Technology; 2001.

Schaefer K. The perception and measurement of human-robot trust [doctoral dissertation]. [Orlando (FL)]: University of Central Florida; 2013.

Schaefer K, Chen J, Szalma J, Hancock P. A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. Human Factors. 2016;58(3):377.

Schmittmann VD, Cramer AO, Waldorp LJ, Epskamp S, Kievit RA, Borsboom D. Deconstructing the construct: a network perspective on psychological phenomena. New Ideas Psychology. 2013;31:43–53.

Segall DO. Development and evaluation of the 1997 ASVAB score scale. Seaside (CA): Defense Manpower Data Center (US); 2004. Report. No.: 2004-002.

Snow RE, Swanson J. Instructional psychology: aptitude, adaptation, and assessment. Annual Review of Psychology. 1992;43(1):583–626.

Stowers K. Communication between human factors psychologists and engineers: challenges and solutions. Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2015. Los Angeles (CA): SAGE Publications. 2015;59(1):1732–1735.

Streater J, Bockelman-Morrow P, Fiore S. Making things that understand people: the beginnings of an interdisciplinary approach for engineering computational social intelligence. Proceedings of the 56th Annual Meeting of the Human Factors and Ergonomics Society Meeting; 2012 Oct 22–26; Boston (MA).

US Army Research Laboratory. The Army Research Laboratory human sciences campaign plan. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2014.

Waytz A, Heafner J, Epley N. The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. Journal of Experimental Social Psychology. 2014;52:113–117.

Wiltshire TJ, Barber D, Fiore SM. Towards modeling social-cognitive mechanisms in robots to facilitate human-robot teaming. Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2013 Sep 30–Oct 4; San Diego, CA.

Appendix A. Principles for Assessment

1. *Assessment must be grounded in a clear purpose*

- Description: Assessments can be used for multiple purposes. As outlined, assessments can be used for selection, determining overall readiness, investigating training or program effectiveness, technology integration, human augmentation, changes to doctrine, or changes to tactics, techniques, and procedures. In addition, assessments can provide formative (e.g., to give feedback or guide further development) or summative feedback (e.g., to determine mastery or impact).

    - Best practice: Given a wide range of potential uses, assessment programs must have a clearly defined purpose to guide development, use, and interpretation of the findings, leveraging theory wherever possible.

        - ✓ *Example*: Assessments designed for one purpose may not be valid for another. For instance, an assessment designed to determine whether an improved targeting device (e.g., a perceptual aid for pilots) may or may not provide information to guide further development (e.g., probability of kill may be insufficient to inform developers where to look for further improvements).

        - ✓ *Example*: To guide development of assessments and interpretation of findings, it is useful to employ a well-articulated, validated theory or model of the proposed augmentation. For instance, a theory of teamwork can guide assessment of planning tools in a tactical operations center by identifying which teamwork behaviors to assess.

    - Best practice: Assessment that serves trainees, operators, and their leaders should support diagnosis, evaluation, and prognosis of knowledge, skills, abilities, behaviors, and beliefs from the schoolhouse to the operational force.

        - ✓ *Example*: Standardized measures should be used whenever possible to identify trends related to a Soldier's performance in the schoolhouse, throughout unit training, and eventually to predeployment training, with the goal of identifying and remediating performance deficiencies early.

    - Best practice: Assessment should support analyses of Doctrine, Organization, Training, Materiel, Leadership and Education, Personnel, and Facilities, part of the Joint Capabilities Integration and Development System process) by providing contextual data concerning those factors to analysts.

- ✓ *Example*: A head-mounted display (HMD) might be designed to enhance Soldier situational awareness. Assessment can be defined to extend beyond this issue. It could address key questions such as: Under what tactical conditions should the HMD not be used? What are the training requirements to effectively use the HMD? What are the logistic requirements for supply?

2. *Assessment must align with the Army's Mission, Values, and Warrior Ethos*

- Description: The Army's Mission has and will continue to evolve given fluidity in its objectives, the nature of adversaries, operational environments, and variations in coalition partners. Currently, emphasis is on learning to thrive and win in the face of complexity, with associated warfighting necessities[1] and human dimension challenges[2] that emphasize 21st-century skills.[3,4] Assessment must support these emerging needs. However, in addition to mission, assessments should reflect and reinforce the Army's cultural values. Culture refers to a pattern of assumptions that are developed by a group over time as they deal with internal and external challenges. These assumptions have been found to work well in the past, so they are considered "correct" and passed on to future members via socialization.[5] Assessments will foster the Army's invariant, core beliefs that transcend time and mission if they are anchored in Army Values (Loyalty, Duty, Respect, Selfless Service, Honor, Integrity, and Personal Courage) and Warrior Ethos ("I will always place the mission first, I will never accept defeat, I will never quit, I will never leave a fallen comrade.").

- The object of assessment must be directly traceable to key elements of the Army's Mission, Values, and Warrior Ethos.

[1] Department of the Army (DA). The US Army operating concept: win in a complex world. Ft Eustis (VA): Training and Doctrine Command (TRADOC); 2014a. TRADOC Pamphlet No.: 525-3-1.

[2] Department of the Army (DA). The US Army human dimension concept. Ft Eustis (VA): Training and Doctrine Command (TRADOC); 2014b. TRADOC Pamphlet No.: 525-3-7.

[3] Department of the Army (DA). The US Army learning concept for 2015. Ft Monroe (VA): Training and Doctrine Command (TRADOC); 2011. TRADOC Pamphlet No.: 525-3-0.

[4] National Research Council. Assessing 21st century skills. Washington (DC): National Academies Press; 2011.

[5] Schein E. Organizational culture. American Psychologist. 1990;45:109–119.

- Best practice: Assessment of technology insertion and Soldier knowledge, skills, and abilities must be traceable to Army Values and mission requirements.
    - ✓ *Example*: Training and augmenting technologies must facilitate the abilities of Soldiers and their units to win in a complex world by maximizing their cognitive, social, and physical potential.
    - ✓ *Example*: Technology insertion must improve mission effectiveness in a manner that is consistent with the Army Values. For instance, advances in technology to promote lethality must not compromise the value of Integrity, defined as the Soldier's ability to do what is legally and morally right. This might manifest itself in evaluating both lethality and potential for collateral damage.

- The process of assessment must be consistent with the Mission and Values of the Army.

- Description: The manner in which the assessment is conducted must also align with the Army Mission and Values. For example, an assessment that has an adverse impact against protected subgroups may not be consistent with Values of Respect (defined as treating others as they should be treated) and Integrity (defined as doing what is legally and morally right), unless it is clearly job- or mission-essential.
    - Best practice: Every measurement method, by itself, is necessarily incomplete and/or potentially biased. Therefore, effective assessment relies on triangulation to cancel out the biases of individual measures, thereby providing a more accurate, holistic, and unbiased assessment of the agent.
    - Best practice: Whenever possible, assessments should be documented and made available to the Army research community at large, so that they can be used to develop new knowledge about human capabilities and limitations, as well as to inspire or evaluate new performance-enhancing technologies, training strategies, and talent management programs.
        - ✓ *Example*: Longitudinal assessment data that tracks the impact of a pharmacological aid or augmentation (e.g., neural stimulation) over time should be widely shared to promote general understanding of Soldier impact, consistent with the Army

Value of Duty (defined as fulfilling one's obligation to accomplish tasks as part of a team).

3. *Assessment should serve multiple users, uses, and time frames.*

- Description: The Army Mission and Values are multidimensional, meaning that one must take measures on multiple aspects of activity to understand it deeply.

- Assessment must be multidimensional.

  - Best practice: Multidimensional assessment enables the users the capability to make tradeoffs and detect potential nth-order effects.

    - ✓ *Example*: Measures of overall effectiveness (e.g., the probability of kill associated with a new squad-based weapon system) should be complemented by or linked to measures of the effects of increased load, enabling understanding of tradeoffs in mobility and lethality.

    - ✓ *Example*: Assessments of a robotic exoskeleton might benefit from measures of cognition and perception that influence Soldier judgment and decision making. Whenever possible, assessments should focus on assessing the Soldier's overall performance (cognitively, socially, and physically).

- Assessment must be multimodal.

  - Best practice: Because the states, behaviors, and effects of agents often cannot be interpreted from a single perspective, assessment must be multimodal. Specifically, assessment should incorporate multiple data sources, measures, and assessments including qualitative sources of data where possible.

    - ✓ *Example*: Assessments taken during a simulated interaction with local civilians involving a new translation tool should incorporate Soldier self-reports, observations from observer controllers, and ratings from the standardized role players.

    - ✓ *Example*: Subject matter expert observations, although potentially qualitative, should and can be included to guide interpretation of quantitative data (e.g., "The Commander's guidance was enhanced given the new decision aid"), although these observations should be explored for reliability and validity wherever possible.

4. *Assessment must produce information whose value exceeds its cost.*

- Best practice: Assessment must generate information that has value in a timely manner. It must produce information that is actionable, meaning that it informs the monitoring and management of events in training, operations, and other domains, and it must do so when there is still sufficient time for corrective action to be taken. Thus, assessment should be formative (not just summative) and continuous (not just discrete) when monitoring and management requirements dictate.

- Best practice: Assessment should be designed to support future audits of assessments and the assessment processes.

    - Example: A unit leader needs to report the readiness of his or her unit. The assessment system must persist records that demonstrate the ways in which that unit is (and is not) mission ready.

- Best practice: Assessment must produce high return on investment and must not divert unduly large amounts of resources from the very activities (e.g., training and operations) that it is designed to benefit.

    - *Example*: Intrusive assessment activities, such as interrupting training to administer inexpensive surveys, may degrade training exercises or operations.

    - *Example*: Unobtrusive sensors used to evaluate the introduction of a new Soldier intervention (e.g., a pharmacological aid), might have high initial costs but might also reduce assessment costs over the long term as data collection is made more efficient relative to observer-based measures.

- Best practice: Assessment systems should employ the best practices and principles of systems engineering to manage costs. These practices and principles include conformance to standards and requirements, interoperability, security, scalability, modularity, and extensibility (which may, in turn, impose requirements concerning intellectual property rights).

- Best practice: The design, development, and operation of assessments should employ scientific best practices (e.g., validation studies) to ensure that the assessments in fact deliver the information they advertise.

    - *Example*: Observer assessments related to the effects of the introduction of a new networked information system for mission planning should be tested for inter-rater reliability.
    - *Example*: Assessments of abstract skills such as leadership should be tested for content and construct validity.
- Best practice: Assessments must be feasible and sustainable in the target environment. They must be usable, meaning that they must be feasible in realistic settings and sustainable in those settings over the long term.
- Best practice: Assessment products must include instruction that enables users to take, interpret, or apply assessments in the manner for which they were designed.

INTENTIONALLY LEFT BLANK.

Appendix B. Lexicon of Assessment-Related Terms (Excerpt)

Word	Definition	Citation
Assessment	Generated by collecting data about the agent, applying a formula (i.e., a measure) to summarize that data and then comparing the summary score (the measurement) with an established performance standard. The data can come from a variety of sources, such as self-reports, expert observations, tools and technologies (e.g., simulators and radio networks), and sensors in the environment.	Plotnik R, Kouyoumdjian H. Introduction to psychology. 10th ed. Thousand Oaks (CA): Cengage Publishing; 2016.
Baseline	A benchmark that is used as a foundation for measuring or comparing current and past performance of a person or system.	VandenBos GR, editor. APA dictionary of psychology. Washington (DC): American Psychological Association; 2007.
Baseline (alternative)	A definition of the functionality, capability, and performance requirements of a system and its interface characteristics.	Ralston A, Reilly E, Hemmendinger D. Encyclopedia of computer science. 4th ed. Hoboken (NJ): Wiley-Blackwell Publishing; 2003.
Capabilities-based assessment (CBA)	The "analysis" portion of the Joint Capabilities Integration and Development System process, in which capability needs are systematically defined. A CBA consists of 6 elements: Scenarios, Functions, Types of Solutions, Capabilities, Concept of Operations, and Measures of Effectiveness. The process answers several key questions for the validation authority including the following: defining the mission; identifying capabilities required; determining the attributes and standards of the capabilities; identifying gaps and their associated risks; prioritizing the gaps; identifying and assessing potential nonmaterial solutions; and providing recommendations for addressing the gaps.	Department of the Army, Training and Doctrine Command (TRADOC). The US Army learning concept for 2015. Ft Eustis (VA): TRADOC; 2012. TRADOC Pamphlet No.: 525-8-2. http://www.tradoc.army.mil/tpubs/pams/tp525-8-2.pdf.
Central tendency	Way of describing the "typical" value in a distribution. Typical measures of central tendency include the arithmetical average of the distribution (mean), the value at the midpoint of the distribution (median), or the most frequent value (mode).	Hays WL. Cohen J. A power primer. Psychological Bulletin. 1992;112(1):155.
Cognitive task analysis	A research method that uses a variety of interview and observation strategies to capture a description of the explicit and implicit knowledge that experts use to perform complex tasks. The captured knowledge is most often transferred to training or the development of expert systems. The outcome is most often a description of the performance objectives, equipment, conceptual knowledge, procedural knowledge, and performance standards used by experts as they perform a task.	Clark RE, Feldon D, van Merriënboer JJ, Yates K, Early S. Cognitive task analysis. In: Spector M, Merrill MD, Elen J, Bishop MJ, editors. Handbook of research on educational communications and technology. Berlin (Germany): Springer; 2008. p. 577–593.
Competency	A cluster of related knowledge, skills, and abilities that affect a major part of an individual's job (a role or responsibility). Correlates with performance on the job and can be measured against accepted standards and improved via training and development. (See Knowledge, Skills, Abilities, and Other Attributes in the document cited.)	Department of the Army, Training and Doctrine Command (TRADOC). The US Army learning concept for 2015. Ft Eustis (VA): TRADOC; 2012. TRADOC Pamphlet No.: 525-8-2. http://www.tradoc.army.mil/tpubs/pams/tp525-8-2.pdf.
Content validity	Extent to which a measure represents all facets of a given target construct. For example, a vigilance scale that focuses only on cognitive dimensions of vigilance, and not on the motivational or affective components of vigilance, would lack evidence for content validity. Like face validity, there is a subjective component to content validity in that all possible dimensions of a construct such as vigilance may not be readily agreed upon.	Coaley K. An introduction to psychological assessment and psychometrics. 2nd ed. Thousand Oaks (CA): Sage Publishing; 2014.

Word	Definition	Citation
Construct	Per Heimen (2002), “an abstract concept used in a particular theoretical manner to relate different behaviors according to their underlying features or causes”. This can include a number of behaviors; for instance, the construct “Resilience” is marked by a wide variety of behaviors, joined by the unifying feature of continuing in light of adversity.	Heiman GW. Research methods in psychology. 3rd ed. Boston (MA): Houghton Mifflin Company; 2002.
Construct validity	Degree to which a measure assesses the theoretical construct it is designed to measure but also *does not* assess what it is *not* designed to measure. Evidence for construct validity can be gathered in 2 ways: 1) by administering existing measures of the construct or related constructs along with the experimental measure and determining if there is a relationship (see convergent validity) or 2) including measures that are supposed to be unrelated to the target construct and demonstrating that no meaningful relationships exist (see divergent validity).	Trochim W. Research methods knowledge base. 3rd ed. Mason (OH): Atomic Dog Publishing; 2006.
Control	Efforts to ensure that physical and situational factors are kept as constant as possible during the course of an experiment for the purposes of isolating the effects of variables of interest.	Hays WL. Statistics. 5th ed. New York (NY): Holt, Rinehart and Winston; 1994.
Convergent validity	Refers to the degree to which a measure or construct is similar to (converges on) another construct that it theoretically should be similar to. Evidence for convergent validity is gathered by administering measures of similar constructs with the target construct and looking for relationships using correlational statistics.	Trochim W. Research methods knowledge base. 3rd ed. Mason (OH): Atomic Dog Publishing; 2006.
Criterion validity	Degree to which a target construct can accurately predict specific indicators of the construct in the real world. Evidence for criterion validity is often gathered by determining the statistical relationship between the target construct and an outcome variable, such as task performance. Either the experimental measure or a previously validated measure are assessed in tandem (see concurrent validity) or the measures are assessed on their ability to forecast future performance later in time (see predictive validity).	Trochim W. Research methods knowledge base. 3rd ed. Mason (OH): Atomic Dog Publishing; 2006.
Dependent variable	A measurable outcome of interest that is not under the experimenter's control but is hypothesized to be affected by independent variables. Because the term “dependent” is meant to describe the outcome “depending” on the independent variable, during nonexperimental designs such as regression analysis, dependent variables should be referred to as criterion variables.	VandenBos GR, editor. APA dictionary of psychology. Washington, DC: American Psychological Association (APA); 2007.
Discriminant validity	The degree to which one construct is dissimilar to or diverges from another construct that it theoretically should not be similar to. By including measures that are supposed to be unrelated to the target construct, evidence for discriminant validity is built by uncovering very weak or nonexistent relationships between measures intended to represent different constructs.	Trochim W. Research methods knowledge base. 3rd ed. Mason (OH): Atomic Dog Publishing; 2006.
Distributed interactive simulation (DIS)	Developed by the Simulation Interoperability Standards Group and approved by The Institute of Electrical and Electronics Engineers, DIS is a network protocol. It describes the exact layout of a few dozen protocol data units that contain information about electronic warfare, logistics, collisions, and simulation management.	McCall M, Murray B. Distributed interactive simulation; 2010 May 10 [accessed 2017 Oct 50]. https://www .sisostds.org/DesktopModules/Bring2 mind/DMX/Download.aspx?Comman d=Core_Download&EntryId=29289 &PortalId=0&TabId=105.

Word	Definition	Citation
Dimension	A continuum on which an individual can have various levels of a characteristic or competency, in contrast to the dichotomous categorical approach in which an individual does or does not possess a characteristic.	Riggio RE. Introduction to industrial organizational psychology. 5th ed. Upper Saddle River (NJ): Pearson Education; 2008.
Examination	A formal test of a person's knowledge or proficiency in a particular subject or skill based on the achievement of objectives.	Riggio RE. Introduction to industrial organizational psychology. 5th ed. Upper Saddle River (NJ): Pearson Education; 2008.
Fidelity	The degree to which a model or simulation reproduces the state and behavior of a real world object, feature or condition. Cognitive fidelity refers to the extent to which a simulator represents the tasks and actions required to train the learning objectives, whereas physical fidelity refers to the extent to which the simulator recreates the sensory components of the real environment.	Hays RT, Singer MJ. Simulation fidelity in training system design: bridging the gap between reality and training. New York (NY): Springer-Verlag; 1989.

Appendix C. The Conceptual Assessment Model (CAM)

The conceptual assessment model (CAM) articulates the finite observables that are used in an assessment, thereby both limiting the scope of the assessment and enabling a clear understanding of all the factors in the assessment (Fig. C-1).

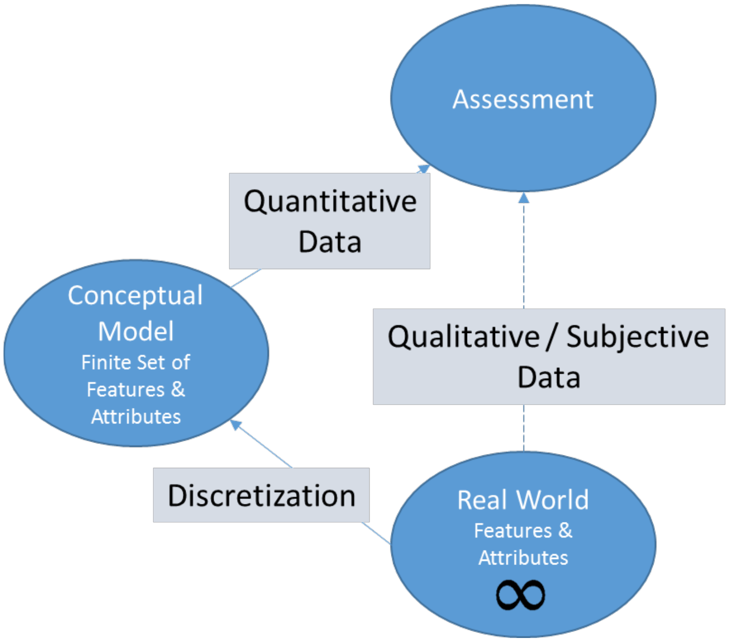


Fig. C-1 Purpose of the conceptual model in Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness (UMMPIREE)

The real world has many features and attributes that may be of interest to a given assessment, so many that the number may approach infinity. The purpose of the conceptual model is to identify and make explicit a finite set of those features and attributes in a way that allows that finite set to be measured. The resulting quantitative data then forms a significant portion of the overall assessment. We recognize that quantitative data alone is not necessarily sufficient for a good assessment and that qualitative and subjective data form key contributions to assessment as well.

Unified Modeling Language (UML) CAM Diagrams

UML diagrams are used as a convenient way of articulating a model structure in a conventionally accepted way. In other words, UML is a commonly used modeling technique. We have here only 2 concepts from UML: classes and compositions. The classes are represented as boxes. Compositions are indicated by diamonds.

Figure C-2 illustrates the CAM using a UML representation. The CAM is composed of one Subject Model, one to several/many Operationalized Constructs, and is associated with one Mission or Assessment Context. This Mission or

Assessment ontext may also influence the Operationalized Construct that is part of the CAM.

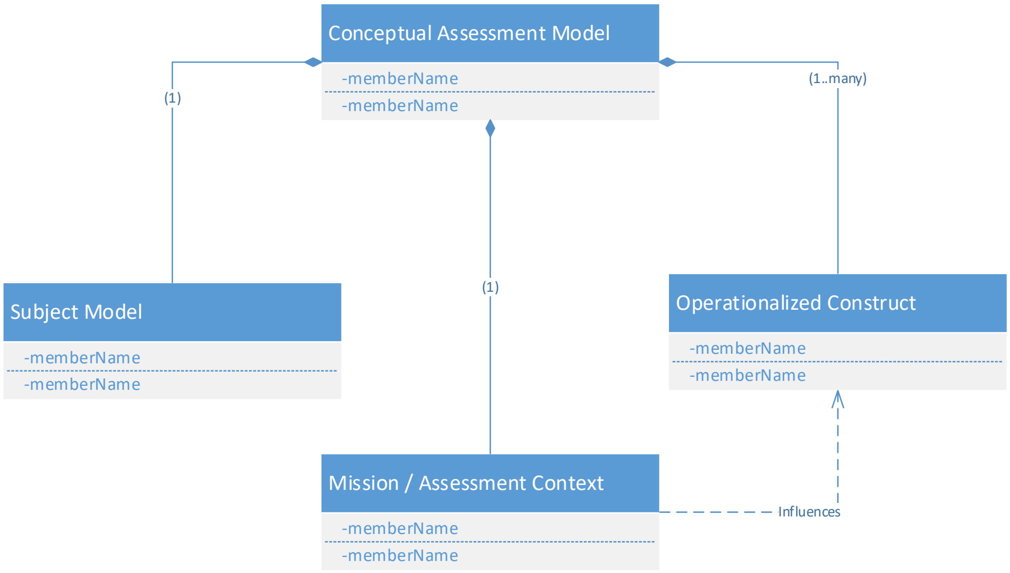


Fig. C-2 Conceptual Assessment Model

The assumption is that "what is being assessed" is the Operationalized Construct, of which there is at least one but could be several. The purpose of the CAM is not to prescribe any particular method of executing assessment (or experiment) but to increase the level of uniformity across similar assessments by framing the assessment in a common, yet flexible structure.

Figure C-3 illustrates the Subject Model class of the CAM. In this example, the Subject Model is specific to the human–machine teaming (HMT) problem space.

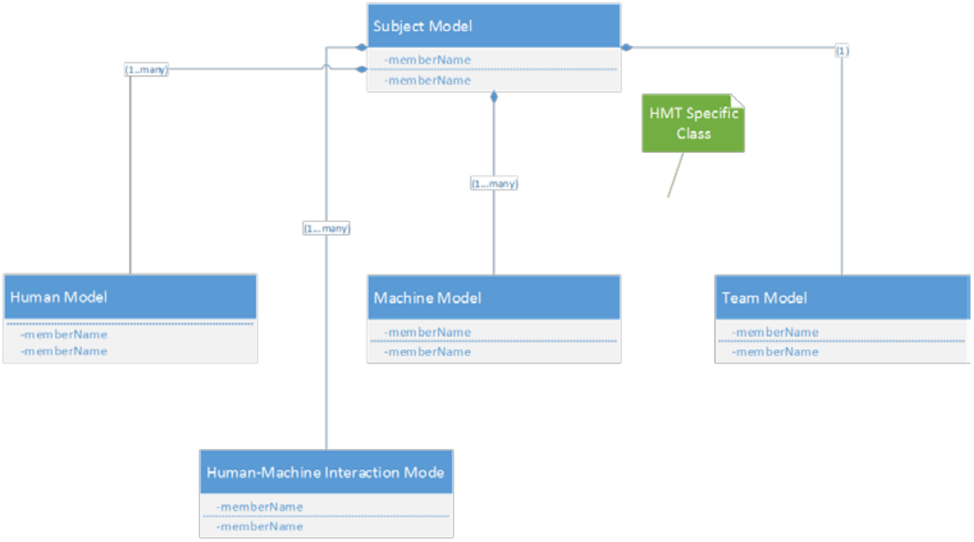


Fig. C-3 CAM Subject Model

Approved for public release; distribution is unlimited.

The Subject Model for the HMT problem space is composed of one to several/many Human Models, one to several/many Machine Models, and one Team Model. In addition, there are one to several/many Human–Machine Interaction modes (e.g., different levels of automation).

Figure C-4 illustrates the Mission or Assessment Context that is an essential element of the CAM and may influence the Operationalized Constructs that compose the CAM.

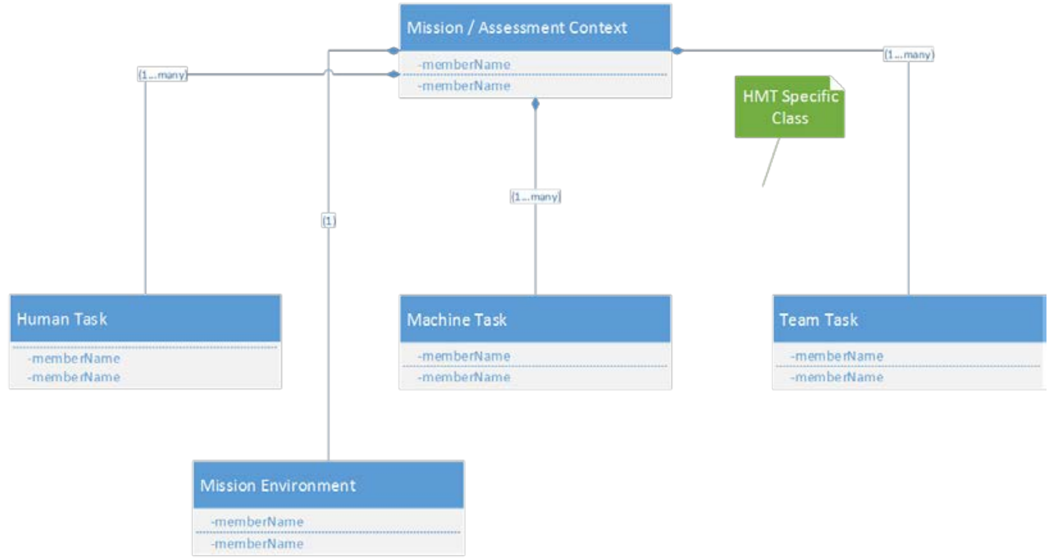


Fig. C-4 The CAM Mission/Assessment Context

This Mission or Assessment Context model is also specific to the HMT problem domain. It is composed of one to several/many Human Tasks, one to several/many Machine Tasks, one to several/many Team Tasks, and a unique (one) Mission (or Assessment) environment.

We define the Operationalized Construct class as shown in Fig. C-5.

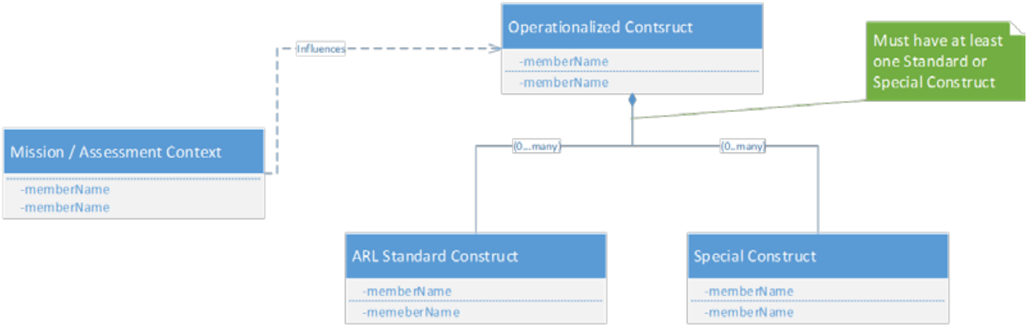


Fig. C-5 CAM Operationalized Construct

The Operationalized Construct is composed of at least one US Army Research Laboratory (ARL) Standard Construct or Special Construct, but there could be multiples of each of these standard and special models. The Operationalized Construct is influenced by the Mission or Assessment Context Model.

We define the Construct Model itself as shown in Fig. C-6. Not surprisingly, the Construct Model can be complex. It can comprise multiple theories, although none are required. The only requirement is that an Evidence Model is defined.

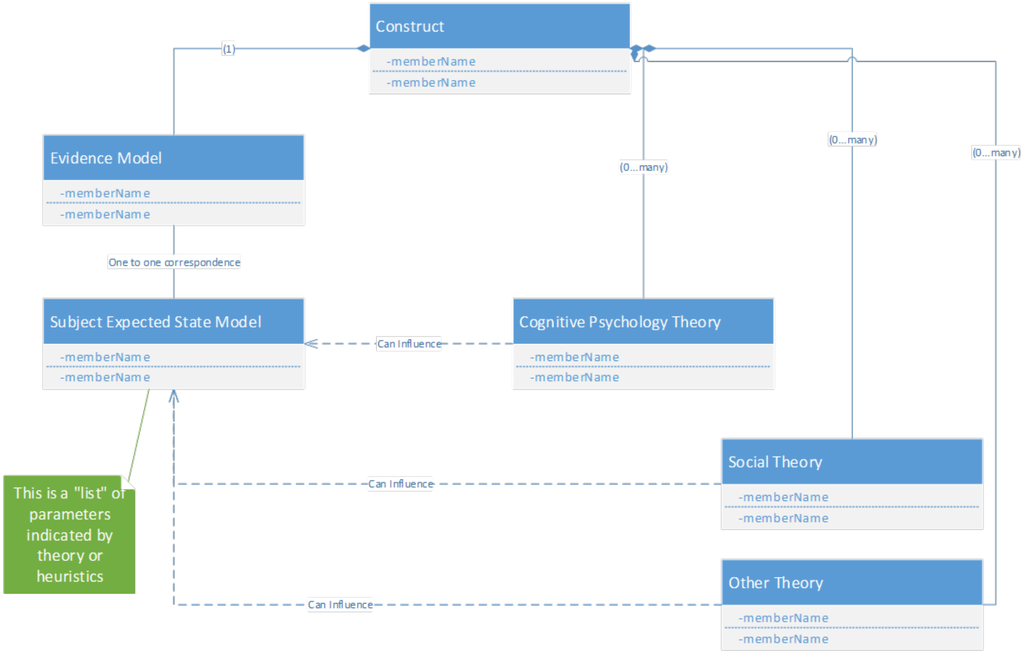


Fig. C-6 CAM Construct Model

In summary, the assessment process can benefit from the use of the CAM through the following steps:

1) Identify and detail (provide specificity) the components of the CAM that will be used for a particular assessment.

2) Develop a data collection and measurement plan for each element of the CAM that is identified as useful for the assessment.

3) Articulate how the components and elements relate to one another from an analysis perspective (How do the tasks relate to the constructs? What data elements will be used for calculating what assessment measures?).

4) Articulate how these data will be analyzed using measures of performance and measures of effectiveness and other high-level measures.

Example of CAM Application

To explore how the CAM might help develop an explicit definition of trust in a specific context, we imagine a simple fictional assessment use case. The situation that we look to assess "trust" in the context of a single Soldier and a robotic mule designed to follow the Soldier while carrying a given load.

If this were an actual assessment (or experiment), we would need to determine how we were going to conduct the assessment, what data we would need, and what measures or analysis would need to be observed and calculated. In this hypothetical example, we simply identify some obvious, and presumably "easy" measures; such measures are highlighted in the following tables.

These tables represent instantiations of the UML classes described previously. Table C-1 includes the particular CAM name (Trust in Soldier–robot teams). The Subject Model is the Soldier–robot. The Mission/Assessment Context name is Transport Heavy Load/Field Environment. We identify 2 operationalized constructs: Trust–will follow and Transparency–Soldier knows state of robot.

Table C-1 Example trust CAM

Name	Name
CAM	Trust in Soldier–robot teams
Subject model	Soldier–robot
Mission/assessment context	Transport heavy load/field environment
Operationalized construct–1	Trust: will follow
Operationalized construct–2	Transparency: Soldier knows state of robot

In Table C-2, our hypothetical example is further developed by describing the Trust in Soldier–robot teams Subject Model. For this table and subsequent tables, several columns are added. These can be thought of as "attributes" of the model. If there are measurable quantities associated with a particular class, those are identified along with suitable measures. For example, the Human–machine interaction mode–1 is a wireless controller. The variable is connectivity and is measured by the percentage of time connected. The final column includes other constraints or characterizations that should be associated with a given class.

Table C-2 Example subject model

Model	Name	Variables/values	Measure	Constraints and characterizations
Subject	Soldier–robot	None	None	None
Human	Soldier	None	None	None
Machine	Robot	None	None	None
Team	Soldier–robot	None	None	None
Human–machine interaction mode–1	Wireless controller	Connectivity	% time connected	None
Human–machine interaction mode–2	Visual	Line of sight (LOS)	% LOS in place	None

Table C-3 describes the example mission-assessment environment. It comprises 2 human tasks, 2 machine tasks, and 1 team task. The mission-assessment environment is an open field, in this case a parking lot.

Table C-3 Example mission-assessment environment

Model	Name	Variables/values	Measure	Constraints and characterizations
Mission-assessment	Go to waypoint in open field	None	None	None
Human task–1	Go from waypoint A to B	None	None	None
Human task–2	Monitor robot	…	…	…
Machine task–1	Follow and maintain pace with Soldier	None	None	None
Machine task–2	Carry load with no damage	…	…	…
Team task	Collaboratively move from waypoint A to B	None	None	None
Mission environment	Open field	None	None	Parking lot

Table C-4 describes the top level of the hypothetical Operationalized Trust Construct.

Table C-4 Example operationalized trust construct (top level)

Model	Name	Variables/values	Measure	Constraints and characterizations
Mission-assessment	Go to waypoint in open field	None	None	None
ARL standard construct	ARL basic 2-party trust	None	None	None
Special construct	ARL HMT Trust	None	None	None

Table C-5 describes the hypothetical ARL Basic 2-Party Trust Construct. In this example, the single feature of the Evidence Model is a reliance agreement between the 2 parties. In this case, the reliance agreement is simply a functioning communications device.

Table C-5 Example ARL basic 2-party trust construct

Model	Name	Variables/values	Measure	Constraints and characterizations
Mission-assessment	ARL basic 2-party trust	None	None	None
Evidence	Reliance	Reliance agreement in place	None	Functioning wireless communications link
Subject expected state	None	None	None	None
Cognitive psychology theory	None	None	None	None
Social theory	None	None	None	None
Other theory	None	None	None	None

CAM Summary

The CAM intends to explore how activities such as human or human–machine team assessments may be improved through a more systematic and standardized approach to defining constructs within a given research or assessment context. Using the UML formalism to define a conceptual model leads to many questions about how constructs are defined and relationships between concepts within such a model. Using a UML class is only a beginning for describing some of the static relationships between concepts. UML (or other modeling approaches for that matter) also provide for ways to further delineate static aspects but also dynamic aspects. This could be particularly relevant for a construct such as trust since trust can be expected to vary over time.

The most important future research is an attempt to use the CAM in a real assessment or experimental setting. The “real world” or “in the wild” settings can be expected to introduce many challenges that could easily overwhelm a CAM implementation that is too literal. This in itself is a challenge to any research intended to further systematize the field of human performance assessment, especially in complex, cognition-intensive, and machine-intelligence-augmented situations. To continue to make progress in this increasingly complex operational environment, progress must be made in systemization and standardization.

INTENTIONALLY LEFT BLANK.

Appendix D. Assessment Planning Framework (APF)

PLAN: Refine goals and framework	
Decisions to be made or hypotheses to be tested	There are many different types of decisions that a sponsor may need to make, and each requires very different information to answer it. Therefore, one must ask critical questions, such as *What is the problem statement? What are the decision maker's critical information requirements? What type of decision must be made? How will the assessment findings be used?* The assessment's specific purpose will significantly guide how it is structured and what the final deliverable will look like. Many assessments fall into one of 3 broad categories: formative, benchmark, or summative.[1] Each type of assessment requires different information to satisfy the decision maker's unique needs. In practice, however, multiple assessments may be combined as part of a larger program of study. For example, a benchmark assessment may be used to identify current levels of marksmanship proficiency across units, and the average training cost per Soldier. Then a summative assessment may be used to compare the effectiveness of different candidate training methods using a sample of representative Soldiers performing representative marksmanship tasks. After selecting the one "best" training method that is to be implemented Army-wide, a formative assessment might be performed after rolling out the new training method in 2 or 3 different units. Based on the feedback and lessons learned from those units, the training intervention may be revised before wide-scale deployment throughout the Army. The key takeaway is that unless the assessment team knows the type of decision that the sponsor needs to make, it will be difficult to determine what type of information to collect, how to organize the results, and how to frame the conclusions.
Intervention or agent under consideration	Relevant questions to be asked in this section include the following: *What technology, training, or intervention is to be assessed? What specific human performance outcomes (e.g., perceptual, cognitive, physical, or social) are expected to result from this intervention?* Simply put, these questions focus on the intended purpose of the underlying technology, intervention, agent, or process that is to be assessed. Fortunately, it is extremely rare that an assessment team will be the first to ever assess a completely new technology, intervention, agent, process, or outcome. In essence, many interventions involve "evolutionary" rather than "revolutionary" change. Since there are likely to be numerous prior attempts and reports documenting the effectiveness of those similar attempts—all of which can be identified via the Defense Technical Information Center or scientific databases such as PsychARTICLES—it behooves the assessment team to carefully review the published literature early on. Doing so provides the following 3 major benefits. First, prior research helps to identify relevant measures, techniques, and technologies that can potentially be reused in the current study. To the extent that measures, techniques, and technologies are considered to be "gold standard" in a given domain, they should generally be incorporated into the planned assessment. The failure to do so would make the results difficult to compare with previous assessment efforts. Moreover, they would also likely lead the project sponsor to be skeptical as to why the assessment team ignored concepts that others had considered to be critically important. Second, prior research helps to identify candidate assessment-related limitations and flaws to be overcome in the current assessment. Research reports are valuable sources of insight for assessment teams. They help to identify gaps in the scientific or operational records, and to the extent that they can be overcome in the current assessment, they allow the assessment team to make a potentially meaningful contribution to both research and practice.

[1] Sanders, JR. The program evaluation standards: how to assess evaluations of educational programs. Kalamazoo (MI): The Joint Committee on Standards for Educational Evaluation; 1994.

Third, prior research provides additional data points to be used in the current assessment. Prior studies do not simply suggest theoretical or methodological issues to be explored in the current study, they actually provide source data to incorporate into the analysis.

Relevant questions to be asked in this section include *What is the larger context in which the assessment will occur? What are the external constraints—including politics, the larger military environment, the tasks, and the intended users—on the assessment process?* These issues help the assessment team to anticipate how their work fits into the larger environment so that they can plan accordingly. While these contextual issues should help to inform the assessment process, they should in no way affect the objectivity of the assessment team or bias their conclusions.

Politics. The term "politics" here refers to potentially sensitive issues that are co-occurring in the larger environment. For example, at the time this document was being prepared, Army and Marine Corps Female Engagement Teams had been serving with distinction in Operation Iraqi Freedom and Operation Enduring Freedom for several years. Not surprisingly, the US Department of Defense has recently begun assessing the effects of mixed-gender units on mission-related outcomes. Given these developments, assessment teams should perhaps consider gender-related issues when planning and designing their assessments. For example, if a new technology is being deployed, assessments should perhaps include a sample of both female and male Soldiers. Similarly, conducting a comprehensive assessment of the technology's effect on Soldier performance—including physical, mental, and emotional effects—would provide a more holistic assessment of the technology's effects, and the results would likely not be biased against female Soldiers who tend to be less physically strong than their male counterparts, but who may be superior in other respects.

Context constraints

The Larger Military Environment. At the time this report was being written, the United States had been at war for over a decade. Access to Soldiers for assessments (particularly research studies) might be limited because the Soldiers are preparing for their next deployment. There may be intense pressure—either explicit or implicit—on unit leaders to certify that their units are ready for deployment, even if they are not. With this in mind, assessment teams may want to purposely exclude superiors during assessment-related events so as to not bias the Soldiers' responses one way or the other. At the very least, assessors should promise the participants that wherever possible, they will only summarize aggregate findings and will not release any personally identifiable information.

The Tasks. The world of work is changing, with virtually every task being augmented by technology in some fashion. Assessors need to understand how technological advances may affect the nature of human performance. To the extent that the technology is advancing quickly, the results of their assessment effort may have a relatively short half-life. In such cases, the appropriate focus may not be on specific technologies (since those will likely change), but on specific classes of technology or human augmentation (which may be less resistant to change). The former assessment would likely focus on just one technology type over a newer version, while the latter would likely involve multiple, similar technologies.

Similarly, when designing assessments, assessors should consider how the tasks are performed in the context of the larger team setting. For example, team tasks require the successful coordination of efforts among multiple individuals. Therefore, individual team members must often suboptimize their own performance for the group to be maximally effective. Task-related performance needs to be considered not in isolation, but in relation to other mission-relevant tasks. In short, human performance may not be monotonic, with "more" not always being better.

	The Intended Users. The Soldiers of today differ systematically from their predecessors, and the Soldiers of tomorrow may be very different from those of today. For example, they may be more technically literate than the Soldiers of today or their predecessors. However, they may not necessarily be as physically strong, fast, or resilient. Therefore, assessors need to consider the population of intended users as it exists today, as well as the likely population in the near future. In order for their assessment results to have a sufficiently long “shelf life”, it may be necessary to conduct the assessment with 2 different groups of participants, those who are representative of “today’s Soldiers” and those who are representative of “tomorrow’s Soldiers”, and pay careful attention to areas where the results converge and diverge.
Theory, model, or doctrine	Assessors should propose an underlying theory, model, or doctrine that fully explains the effects of the augmentation on human performance. For example, the theory should answer the following question: *Is the proposed human augmentation a main effect, a mediated effect, or a moderated effect?* Similarly, the theory should help to specify *What constructs should we measure?* and *When should we expect to observe an effect?* In many cases, human performance is multidimensional and includes elements of physical, cognitive, emotional, and/or perceptual skills. If so, those skills should be specifically listed, so that a holistic assessment of the human performance augmentation can be established. The theory should provide some guidance on when the effect should become visible. While some effects may be relatively immediate (e.g., the introduction of a new performance-enhancing drug or technology), others may take time before they are readily apparent (e.g., the introduction of new policies or procedures). If the assessor is measuring the correct constructs but measuring them at the wrong time, they will likely fail to find the proposed effect. A well-developed theory directs the assessor where and when to focus the assessment in order to find the proposed effect.
Metric and baseline or standards	Relevant questions to be asked in this section include the following: *On what scale should human performance be measured? What are the criteria of effectiveness or proficiency? What human performance outcomes are considered practically meaningful?* These questions help the assessment team to accurately quantify human performance as well as changes in human performance that result from augmentations. Virtually every skilled task in the Army has some predefined measurement standard and associated scoring criteria. If assessors want their work to be perceived as credible, they must work closely with Army subject matter experts (SMEs) to ensure that they are measuring human performance according to established Army standards. Even if the assessors decide to include a new scoring or assessment mechanism, they should present the current assessment metric for comparison purposes. Without this, it will be impossible to compare the effects of the current study with prior studies. By extension, when reporting about improvements in human performance—for example, those caused by technological augmentations—assessors should reference the appropriate Army metrics. Depending on the cost and/or availability of the technological augmentation, certain improvements may be considered more practically meaningful than others. The practical meaningfulness of human performance augmentations must be considered in context. Standardized metrics—for example, with regard to performance and/or cost—should be used to help establish practical meaningfulness.
Resource constraints	Every assessment-related effort requires an investment of resources, such as funding, time, access to Soldiers as participants, access to facilities such as training ranges for data collection, and related factors. When planning their efforts, assessment teams need to be realistic about what resources they need to conduct their work, and must recognize that the resources that they actually get may be very different from what they have requested. Therefore, assessment teams need to consider questions such as the following: *How can we work within the likely constraints to design the best possible assessment? How will these constraints impact our findings?* While these resource constraints exert independent effects on the assessment team, their combined effect on the overall assessment design can be greater than the sum of the individual

	parts. For example, issues of funding and time typically go hand in hand, for no other reason than the fact that budgets are made annually. Therefore, even if the best possible assessment would require longitudinal data collection (e.g., 2–3 years), the budget may require that the assessment be completed within 6 months or the funding may become unavailable. As a result, the assessment team may need to scale back the breadth of its intervention, such as by measuring human performance for a much shorter period of time, which may result in smaller sample sizes. If that is the case, perhaps the assessment should include more measures to yield more-detailed information from the smaller sample of Soldiers, since a larger sample may not be attainable. Every assessment team faces constraints. In some cases, the constraints are firm and fixed. In other cases, the constraints may be negotiable. It is generally advisable that the assessment team put together high-level descriptions of different research designs along with their inherent strengths and limitations. Armed with this information, they can approach their sponsors and perhaps make a convincing case for relaxing some of the constraints.
Assessment design	There are a number of possible research designs, the most common ones being the controlled experiment, the quasi-experiment, and the naturalistic observation. While good for testing causal relationships and internal validity, experiments can sometimes be artificial relative to the real world. For example, the experimental laboratory or test bed may provide only limited real-world cues that participants rely on when doing their work-related tasks. Quasi-experimental designs attempt to statistically or methodologically control for the lack of random assignment, with varying degrees of confidence. These designs represent plausible alternative approaches, especially when the advantages of random assignment do not outweigh the disadvantages, such as an unrepresentative task or environment. Naturalistic observation is used when the assessment team is merely measuring behavior or performance but not attempting to systematically control what happens in the larger environment. This approach is often used in benchmark studies, since the goal is to establish how things are performed currently or historically. All 3 approaches have their unique strengths and weaknesses. No one approach is "best" in an absolute sense. It is up to the assessment team to weigh the relative costs and benefits of each approach and then choose the one that best "fits" with all of the other constraints described.
PLAN: Define methods	
Environment	There are a number of factors to consider regarding the assessment environment. One of the most critical is the scenarios that will help to elicit the participants' behavior. Simply put, one does not assess a team's situation awareness or performance in the abstract. These types of characteristics can only be assessed in the context of a mission-relevant task. The key here is for the assessment team to work closely with SMEs to design scenarios that are task and mission relevant. To the extent that the scenarios are not representative of the target of interest, the assessment findings will likely be called into question. After specifying the scenarios, the assessment team will need to identify suitable facilities or the physical location where the assessment team will collect their data. This may be an indoor laboratory or an outdoor training range. Factors to consider when selecting the test facility include task suitability, availability, location, schedule, classification level, and safety considerations, among others. In many cases, there will be more than one test environment that could potentially be used. Since no one environment is likely to be "best" on every single dimension, assessment teams will need to carefully weigh their options.

Participants	There are a number of factors to consider with regard to the assessment participants. Perhaps the most important consideration is the number of participants required to achieve "statistically significant" results, for example, based on a statistical power analysis (Cohen 1992).[1] Generally speaking, the larger the sample size, the greater the ability to detect even small or trivial differences between groups using traditional Null Hypothesis Statistical Test (NHST) methods. However, large sample sizes are very difficult to obtain. Therefore, assessment teams may opt to employ less elaborate research designs, which by extension require fewer participants. Alternatively, the assessment team may seek to use Bayesian statistical estimation methods, which also require fewer participants than the NHST approach. The characteristics of the sample also matter. Assessment teams should generally try to recruit participants who are similar to the intended target audience. However, in many cases, the assessment team is unable to obtain the type of Soldiers with the background or experience needed and has to resort to whoever is available. Therefore, the assessment team should collect detailed background information from each Soldier or team to be able to statistically "control" for extraneous effects using multiple regression, analysis of covariance, or related techniques. Even if the assessment participants closely resemble the intended population on key characteristics, statistical covariates can still be used in the analysis process, if for no other reason than to show that these factors did not affect the results.
Measures	When thinking about potential measures, assessors should be prepared to ask 3 critical questions: *What constructs should we measure? How should we measure these constructs? When should we observe the effects?* There are a number of task- and mission-related constructs that one could potentially measure. Some of these measures are process-based (i.e., behaviors), while others are outcome-based. The assessment team should attempt to measure both processes and outcomes whenever possible. The question of "how to measure" involves a number of considerations, such as the extent to which data are available for measurement, whether the data meet psychometric standards of reliability and validity, and the ease with which data can be collected. In many cases, the desired data may be difficult to collect; for example, if the weapon, tool, or system being used by the Soldier is not instrumented. In such cases, the assessment team may need to rely on costly expert observer-based ratings. Another critical consideration is data reliability and validity. By definition, reliable measures provide highly similar results under highly similar conditions. Similarly, valid measures provide results about the construct of interest and are not systematically contaminated by other factors. Reliability is a necessary component for valid measures. A measure cannot be valid if it is not reliable. Finally, the question of "when to measure" again involves one's underlying theory of human performance. Depending on the specific human performance augmentation, its effects on performance may be immediate or delayed. As noted previously, even if the assessor is measuring the correct constructs—but measuring them at the wrong time—they will likely fail to find the proposed effect.
Procedure	Whenever possible, the assessment team should seek to standardize the data collection process as much as possible to rule out potential confounds. Doing so will help to ensure that the assessment results are readily interpretable and not contaminated by extraneous factors. There are 3 primary ways that the assessment team can standardize the process: controlling extraneous conditions, using a standardized data collection script, and using a standardized process for collecting and handling the data. These 3 mechanisms are typically used together.
	Assessment teams need to explicitly consider how they will document their data. The documentation could occur in various ways, for example by including electronic copies of all measurement instruments in the same folder with the data files, or by including detailed notebooks which describe the specific procedures by which the source data were collected, integrated, cleaned, checked for accuracy, and de-identified (as appropriate). Many times,

[1] Cohen J. A power primer. Psychological Bulletin. 1992;112(1):155.

<table>
<tr><td><b><i>Analysis</i></b></td><td>assessment teams will write “syntax files”—in packages such as R, SPSS, or SAS software—that allow them to easily recreate all of their analyses by simply clicking a few buttons. This approach has several benefits, such as allowing other researchers to replicate your specific analyses. It also allows the research team to annotate its analyses using comment statements.

One of the biggest challenges for assessment teams, researchers, and scientists is the need to reformat data for different uses. Modern data scientists have created an extremely robust approach for storing data such that they are easily interpretable to both humans and computers. Having the data stored in the right format is particularly important to assessment teams, as they may need to share their data with others, or incorporate data that others have collected into their own efforts.</td></tr>
<tr><td colspan="2"><b>Develop materials, pilot test, and report</b></td></tr>
<tr><td colspan="2">The outcome of planning is a detailed roadmap for conducting the assessment proper. When all of the “big picture” decisions have been made, the assessment team can begin developing the assessment-related materials. This may entail designing or buying measurement instruments, developing stimuli, and training the assessment team members who will act as field observers or data analysts. A pilot test is well advised, which may lead the team to revise methods and materials. Next, the team executes the assessment, analyzes the assessment data, and issues a report of findings and recommendations to the sponsor who, in turn, may make a decision or take action that is proportionate to the cost of the assessment. The decision may shape future assessment tasking from that sponsor. Feedback from the sponsors and from assessment team members supports an after-action review, in which the assessment team identifies lessons learned and revises its assessment procedures to improve its future work.</td></tr>
</table>

INTENTIONALLY LEFT BLANK.

Appendix E. Assessment Stakeholder Tiers

Army Stakeholders: Tier 1

Stakeholder	Discussion
Soldier	US Army Soldiers at all echelons and many military occupational specialties will benefit through the use of more scientifically vetted training and operational capabilities, enabling greater mission success. Long term, this benefit is aimed at producing "cognitive readiness" as part of a broader portfolio of force readiness.
Trainer	Those responsible for training Soldiers will benefit by using well-understood training mechanisms whose effects on the Soldier and his or her ability to effectively progress in individual, team, and mixed-team cognitive skills.
Various Army organizations	Various Army organizations will benefit from the Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness (UMMPIREE). They are too numerous to list here, but examples include the following: **Training and Doctrine Command** will benefit from support to doctrine development through assessment of cognitive skills in various training and operational settings. **Army Capabilities Integration Center** will benefit through increased confidence in the integration of new cognitive-based tactics, techniques, and procedures (TTPs) and systems. **Mission Command Battle Lab** will benefit through an increased ability to understand how augmentation may impact both the mission and Soldier's cognitive state while executing missions of the future. **Combatant Commands** will benefit through both improved training and overall better human performance assessments

Management Actors: Tier 2.

These actors are primarily invested in UMMPIREE and its resulting products from a programmatic sense.

Actor	Discussion
Human/team/mixed-agent evaluator/assessor	This stakeholder is typically the lead researcher or principal investigator who may also have a team of cognitive scientists, engineers, and other SMEs at his or her disposal. This stakeholder is interested in using established methods and tools to make an assessment of humans, human teams, or mixed-agent teams in their performance of a particular set of tasks. The subject of the evaluator/assessor's investigation is varied and could include combat development, human performance enhancement, and other areas. This stakeholder is sometimes the same person as other researchers/developers in this Management Stakeholder list. This stakeholder also appears in Tier 3.
Human assessment researcher	The human assessment researcher is interested in understanding how human assessment (or team or mixed-team assessment) is conducted and how to improve or better understand that process.
Operational system developer	The operational system developer is trying to understand the interaction of human/team/mixed-team performance with an operational system (i.e., a combination of hardware, software, and networks) under development.

Actor	Discussion
Training system developer	The training system developer wants to explore the interaction of the trainee with a particular training system under development. The training system developer wants to ensure that the training system will be effective and uses assessment of the trainee as feedback in developing the training system.
Human sensing system developer	The human sensing developer wants to explore the interaction of specific sensors and how well these sensors can be used in an assessment environment and/or an operational environment. "Sensor" is used in a broad sense to include many things from physical devices to questionnaires.
Augmentation system developer	The augmentation system developer wants to explore the interaction of specific augmenting technologies on human (or team or mixed-agent) performance.
Human performance researcher	The human performance researcher is interested in understanding how to understand and improve performance of individuals and teams through selection, training, aiding, organizational design, process design, and other means.
Combat developer	The combat developer is interested in TTPs and doctrine associated with changing operational environments, Army capabilities, and technologies. The combat developer will be interested in how humans and mixed teams perform in new and novel environments that will then impact the TTPs and doctrine development.
US Army Research Laboratory (ARL)	ARL needs an organizational capability to support research and development efforts in assessment of training and augmentation capabilities.

Cognitive Science and Engineering Actors: Tier 3

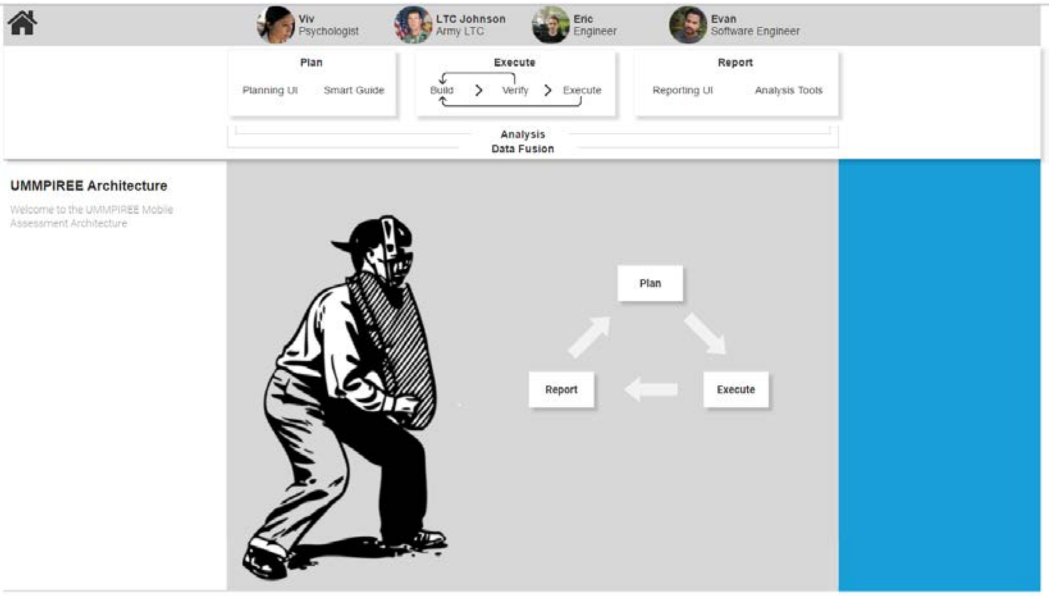
These stakeholders interact most directly with UMMPIREE products (methodology and tools). In many cases, a particular individual will act in the role of several stakeholders. For example, the assessment planner may also be the analyst. The event/experiment controller may also be the event/experiment observer. Not all stakeholders may be used or needed for every event or experiment.

Actor	Discussion
Assessment planner	This stakeholder wants to plan an assessment event. The event could be an experiment or a standard evaluation of individuals, teams, or mixed-agents. The planner will often have an incomplete idea of exactly how to conduct an event and will expect some help from the assessment execution environment infrastructure. The planner will also likely have constraints, limitations, and assumptions that he or she brings to the environment.
Event/experiment observer	This stakeholder wants to observe an event or experiment and may do so with or without specific tools (e.g., software or devices).
Event/experiment controller	This stakeholder needs to control certain aspects of the event or experiment. Things that might need to be under the controller's purview include timing and nature of specific interactions between the human (team or mixed-agent), the system itself, and/or the training or augmentation capability.

Actor	Discussion
Human/team/mixed-agent evaluator/assessor	This stakeholder appears in the second and third stakeholder tiers. In this (third) tier, the evaluator/assessor is involved in the actual conduct of the event through the use of tools (e.g., data collection).
Analyst	The analyst uses the Assessment Execution Environment to examine the collected data, analyze those data against the particular objectives and assessment methodologies in use, and develop analytic conclusions based on the data and methods. Generally, the analyst does not provide the final or ultimate assessment—a task associated with the lead assessor/evaluator—but does provide processed material to allow the lead to make informed assessments/evaluations.
Human/team/mixed-agent behavior modeler	This stakeholder may use the Assessment Execution Environment in one of several modes (to include consultation) to develop or modify (edit) models of human/team/mixed-agent behavior. These models may be textually or mathematically descriptive and/or encoded in software and data.
Human variability and performance data archivist	This stakeholder is responsible for organizing data collected through the use of the Assessment Execution Environment (including "external" data available through the environment) to build and refine models of human variability and performance. These models may be useful in themselves but are also useful in support of events/experiments.
Human assessment methods engineer/scientist	This stakeholder is generally a researcher into methods and techniques (i.e., the science) of assessment. This stakeholder may observe ongoing events/experiments or may use the environment databases in support of specific research questions aimed at improving the science of assessment.
Human assessment sensing engineer	This stakeholder is interested in how a given sensing device or method can be used for assessment. This may involve the conduct of experiments and/or accessing the environment databases for relevant information.

Appendix F. Mobile Assessment Architecture (MAA), Version 2.0

Figure F-1 shows an example of version 2 of the Mobile Assessment Architecture (MAA).



UMMPIREE Architecture

Welcome to the UMMPIREE Mobile Assessment Architecture

Fig. F-1   MAA version 2

Components of the MAA

The MAA architecture (Fig. F-2) consists of 3 main system components, Plan, Execute, and Report, and has 2 underlying subsystems, analysis and data fusion, that interact with all 3 of the main components.

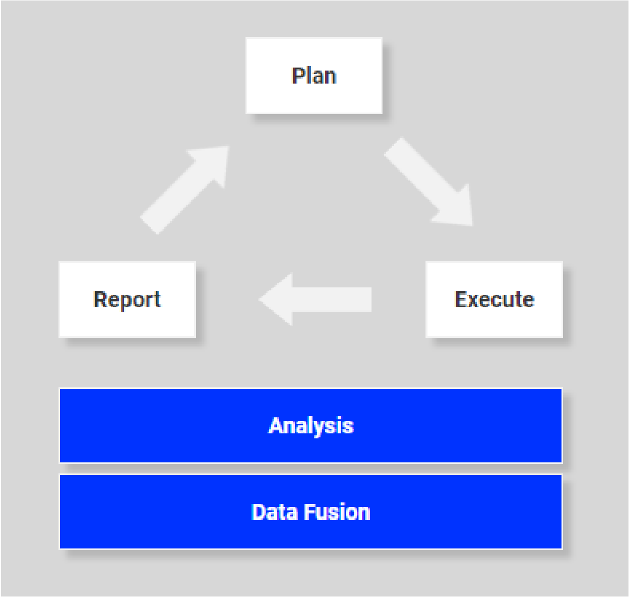


Fig. F-2   MAA systems and subsystems

Approved for public release; distribution is unlimited.

Plan

The Plan (Fig. F-3) component enables users to plan an assessment exercise using the Smart Guide, to view and edit various planning documents and scripts that it generates, and maintain the Tool Box from which the Smart Guide composes assessment solution packages.

The user interacts with the Plan component through a Planning User Interface (UI) that interfaces with a Tool Box, a Smart Guide, and an Assessment Plan.

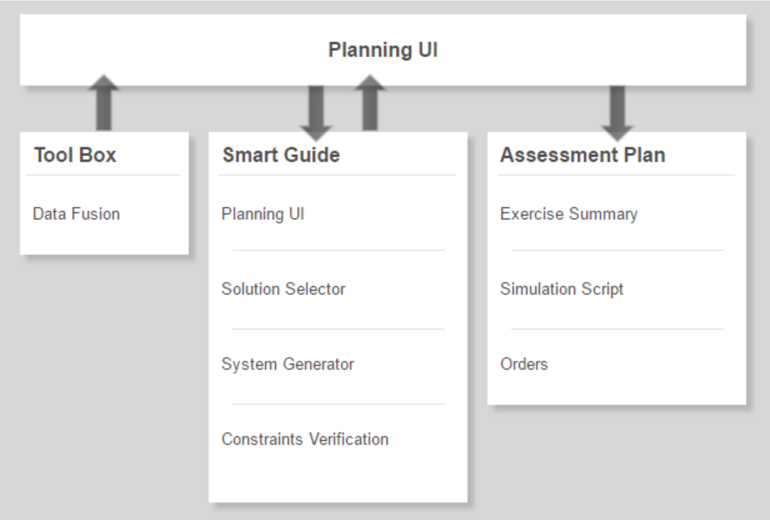


Fig. F-3 Plan component architecture diagram

The Tool Box is a collection of resources available for use in a given assessment exercise, including facilities, participants, staff, instruments, measures, and assessments. These resources might be used for a given assessment exercise along with metadata that support tradeoff analyses. Metadata could include financial costs, time costs, availability, compatibility, and the like. The Smart Guide assists the user in eliciting requirements, identifying constraints, and selecting assessment solutions. It guides the user by presenting questions that will aid the user in determining requirements and constraints for the exercise. It also helps identify tradeoffs within and between solution packages, and presents solution packages with tradeoff guidance for editing by the user. The Assessment Plan provides with the details needed for exercise execution. It consists of several types of subsystem output, including 1) an assessment exercise summary describing exercise objectives, participants, procedures (technologies, scenarios, measures, staff, analysis methods, protocols, etc.), and schedules, 2) orders for certain resources (staff, technology), and 3) a simulation generation script that automatically integrates existing objects into a simulation environment that may be deployed in analysis and in execution.

Execute

The Execute component (Fig. F-4) will build upon the concepts identified in the Plan component. This is where the experiment is built, validated, and tested. This phase is an iterative process where feedback from each component, both preceding and subsequent, is incorporated throughout simulated and live execution.

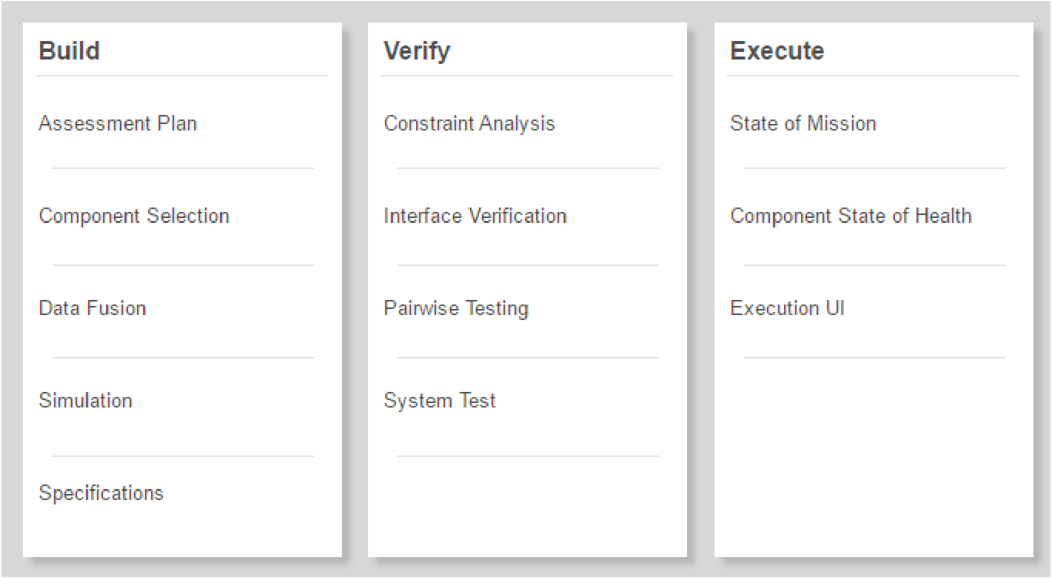


Fig. F-4 Execute component architecture diagram

The Build subcomponent focuses on taking the previously generated assessment plan and assembling it for thorough analysis and testing. Items will be selected and scrutinized, data fusion functions are identified, and the assessment plan endures rigorous simulation and the subsequent specifications are identified. During the Verify phase, the schedule and possible logistical hurdles of the proposed assessment (e.g., a timeline) are addressed to ensure that significant milestones are satisfied. The verification process also serves as an inspection period for the simulations conducted in the previous stages (i.e., Does the plan that was developed and simulated seem to have a high probability of success?). The Execution Map depicts the entire execution process and the roles that would be required for proper implementation and execution of the assessment plan.

Report

The Report component (Fig. F-5) includes a user interface that directs reporting activities and postexperiment inquiries. Some prospective examples include after-action reviews (AARs) and technical/summary report generation. Simply stated, this phase assists the user in understanding any additional analysis needed or recommended action items.

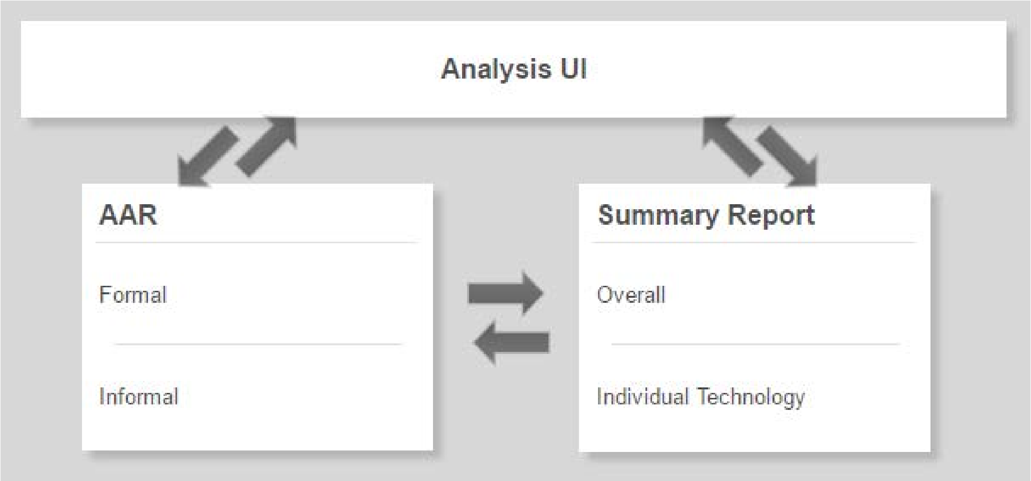


Fig. F-5 Report component architecture diagram

The Analysis UI is presented in real time and post hoc. It also assesses real-time data significance to help the user determine whether he/she needs to continue to collect data or has enough data for an accurate assessment. Content for both the AAR and Summary Report is created and viewable by the user. The AAR is used at the end of the assessment to apprise leaders and others involved of quick lessons learned, what went right, and what could have been improved. The Summary Report consists of a document (e.g., PowerPoint) that summarizes the experiment and findings in such a way that readers can rapidly become acquainted with the study and its outcome from a high-level perspective. It can show information on all technologies utilized or focus on a specific technology used during an assessment.

Subsystems of the MAA

Data Fusion

The Data Fusion subcomponent (Fig. F-6) supports the integration of multiple data and information assets into a consistent, accurate, and useful representation of the system. This integration consists of data inspection, cleansing, and transforming to provide structured and usable data across all components and across multiple assessments. Live, real-time data for actual participants on a physical system can be used as well as simulated data involving actual participants or simulated participants on a simulated system. The primary role of Data Fusion is to provide an extensible layer that can support arbitrary data sources and types within the system's domain. Data Fusion consists of 3 layers: data conditioning, interoperability, and the dynamic management of the raw data.

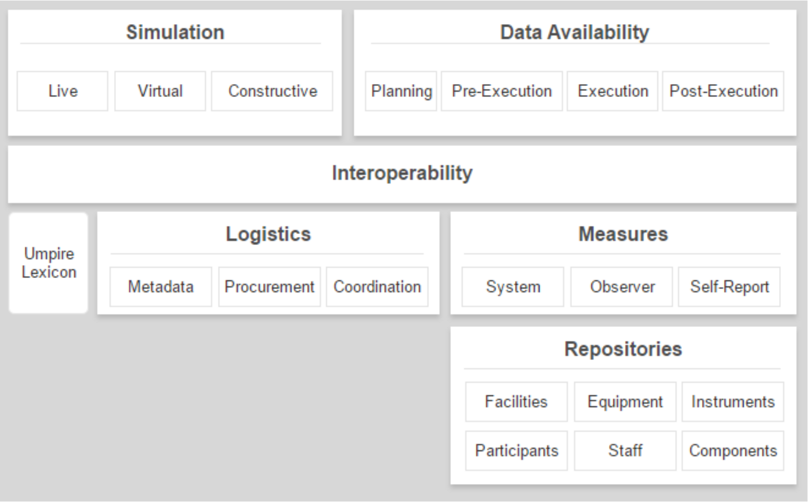


Fig. F-6 Data analysis architecture diagram

Analysis

The Analysis subcomponent (Fig. F-7) consists of real-time assessment management by amending current descriptive statistics with inferential statistics to determine which measures we must continue to collect and which measures we have enough of to make decisions or support hypotheses. User facing tools include analysis tools, a commas-separated values (CSVs) exporter, and statistical package interfaces. The Analysis subcomponent supports discovery of useful information, assists in drawing supported conclusions, and aids in decision making. It blends together the 3 primary system components and can incorporate information from previously run assessments to provide the most complete analysis available.

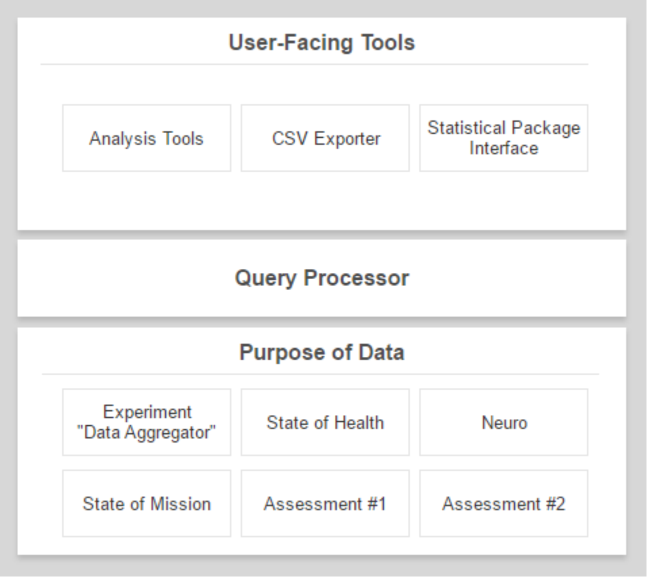
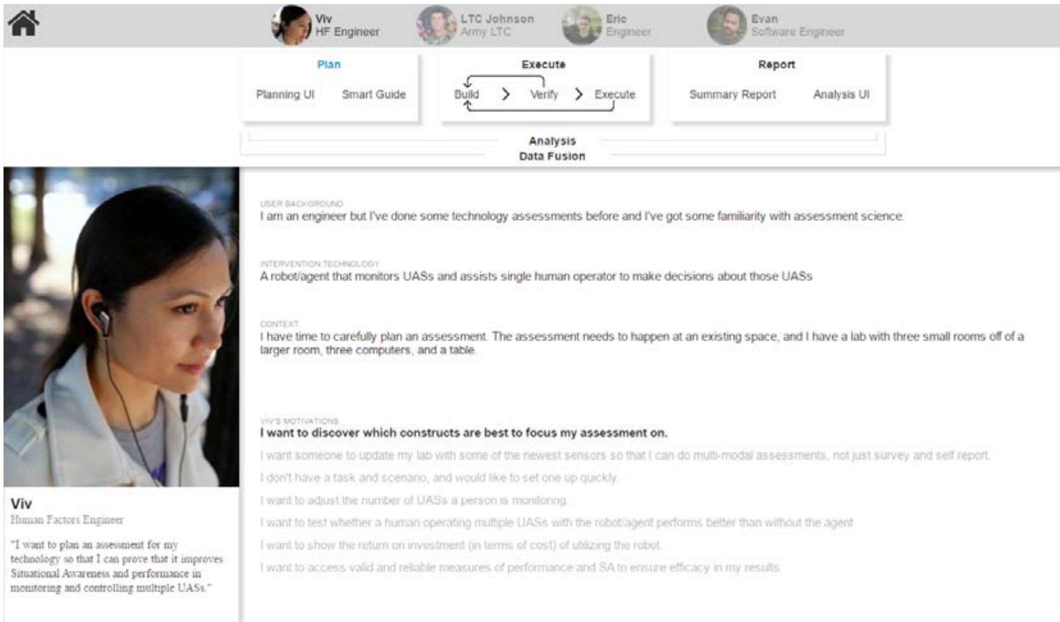


Fig. F-7 Analysis architecture diagram

User Stories

User stories (from the Tier 3 actor category) are included to support current and future MAA development activities. The user stories can be found in the circles at the top of the main page (Fig. F-8). If you click on a user, you will see their background, augmentation technologies, context in which they are using the MAA, as well as specific things they want to be able to do with the MAA. In the example in Fig. F-8, Viv is a Human Factors Engineer. She is building an agent that assists unmanned aircraft system operations with decision making. Viv has some experience with conducting assessments but would like some help with planning her assessment. One specific need is that she would like to discover additional constructs that would enhance her assessment. On this page, the MAA is pointing out that tools meeting that requirement would sit in the "Plan" category of tools.



USER BACKGROUND
I am an engineer but I've done some technology assessments before and I've got some familiarity with assessment science.

INTERVENTION TECHNOLOGY
A robot/agent that monitors UASs and assists single human operator to make decisions about those UASs

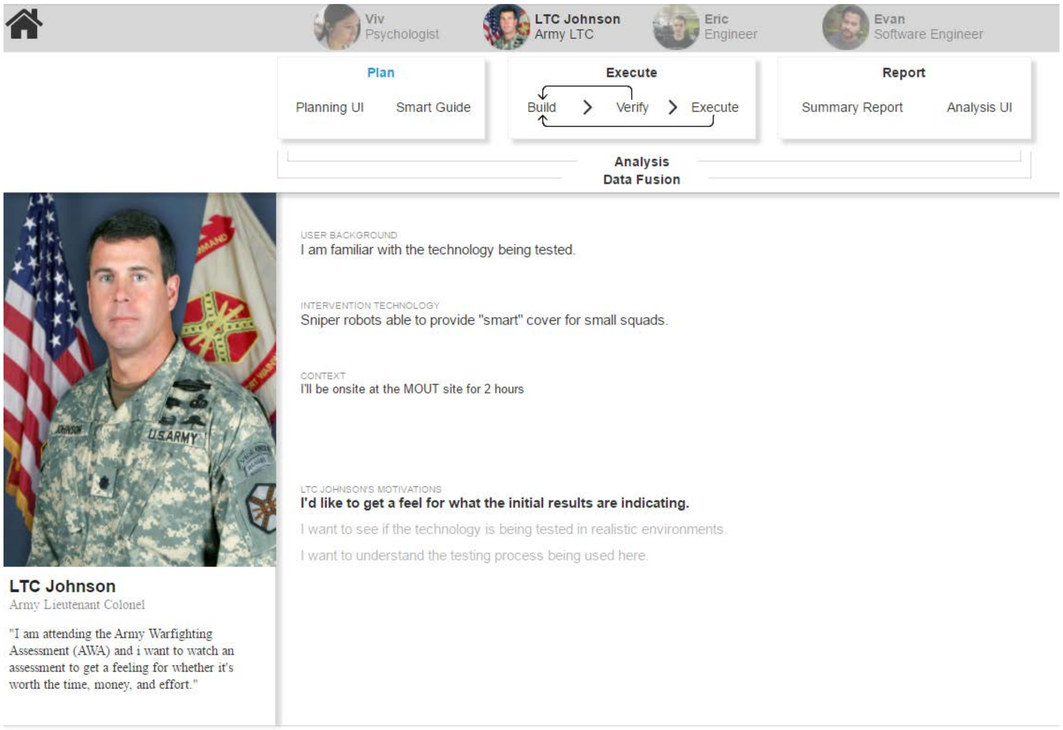
CONTEXT
I have time to carefully plan an assessment. The assessment needs to happen at an existing space, and I have a lab with three small rooms off of a larger room, three computers, and a table.

VIV'S MOTIVATIONS
I want to discover which constructs are best to focus my assessment on.
I want someone to update my lab with some of the newest sensors so that I can do multi-modal assessments, not just survey and self report.
I don't have a task and scenario, and would like to set one up quickly.
I want to adjust the number of UASs a person is monitoring.
I want to test whether a human operating multiple UASs with the robot/agent performs better than without the agent
I want to show the return on investment (in terms of cost) of utilizing the robot.
I want to access valid and reliable measures of performance and SA to ensure efficacy in my results.

Viv
Human Factors Engineer
"I want to plan an assessment for my technology so that I can prove that it improves Situational Awareness and performance in monitoring and controlling multiple UASs."

Fig. F-8 MAA user story (Viv)

In another example (Fig. F-9), Lieutenant Colonel (LTC) Johnson is a stakeholder for a new technology: sniper robots that provide "smart" cover for small squads. He will be onsite for 3 h at a MOUT (military operations on urban terrain) site to observe an official assessment of that technology. During that time, he would like to get a feel for what the initial results are indicating. On this page the MAA is pointing out that tools that would meet that requirement would sit in the "Execute" category of tools.



USER BACKGROUND
I am familiar with the technology being tested.

INTERVENTION TECHNOLOGY
Sniper robots able to provide "smart" cover for small squads.

CONTEXT
I'll be onsite at the MOUT site for 2 hours

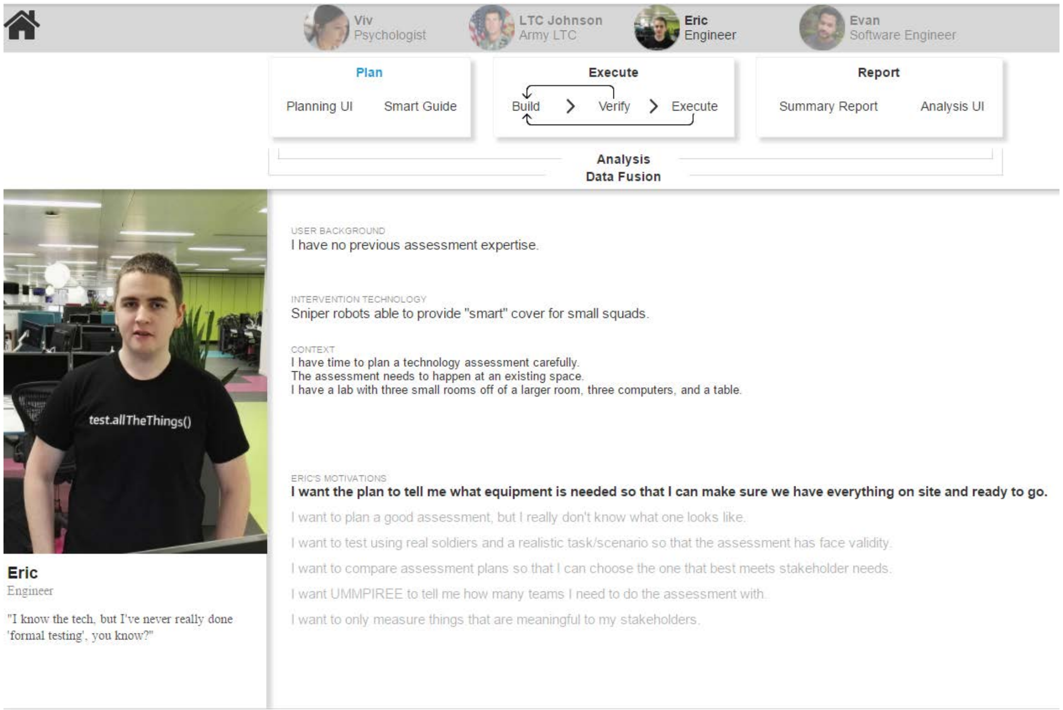
LTC JOHNSON'S MOTIVATIONS
I'd like to get a feel for what the initial results are indicating.
I want to see if the technology is being tested in realistic environments.
I want to understand the testing process being used here.

LTC Johnson
Army Lieutenant Colonel

"I am attending the Army Warfighting Assessment (AWA) and i want to watch an assessment to get a feeling for whether it's worth the time, money, and effort."

Fig. F-9 MAA user story (LTC Johnson)

A third example of a user story is shown in Fig. F-10. Eric is the engineer of these sniper robots that provide “smart” cover for squads. His government stakeholder has asked that he test whether the robot will have any adverse impact on squad performance. Eric has no prior experience in setting up such an assessment. One particular need is that he wants to know what additional equipment he needs to do the assessment given the current space in which he can conduct the assessment. For this user story, the MAA identified that that need will be met in the “Planning” set of tools.



USER BACKGROUND
I have no previous assessment expertise.

INTERVENTION TECHNOLOGY
Sniper robots able to provide "smart" cover for small squads.

CONTEXT
I have time to plan a technology assessment carefully.
The assessment needs to happen at an existing space.
I have a lab with three small rooms off of a larger room, three computers, and a table.

ERIC'S MOTIVATIONS
I want the plan to tell me what equipment is needed so that I can make sure we have everything on site and ready to go.

I want to plan a good assessment, but I really don't know what one looks like.

I want to test using real soldiers and a realistic task/scenario so that the assessment has face validity.

I want to compare assessment plans so that I can choose the one that best meets stakeholder needs.

I want UMMPIREE to tell me how many teams I need to do the assessment with.

I want to only measure things that are meaningful to my stakeholders.

Eric
Engineer

"I know the tech, but I've never really done 'formal testing', you know?"

Fig. F-10 MAA user story (engineer)

Conclusions and Future Integration

Currently, the MAA serves as a high-level architecture under which assessment solutions can be organized and linked. In the near future, we plan to focus on a specific set of tools that will be placed into the MAA architecture.

INTENTIONALLY LEFT BLANK.

Appendix G. Subject Matter Expert (SME) Interview Questions

Section A: What the researcher is currently doing

1) What is your area of interest? (criteria, dependent variables [DVs], effects)

2) Within this area, what do you study? (unit of analysis)

3) Who is the target of your intervention/training? (unit of analysis)

4) How would you assess the DVs/criteria?

    - What assessments are you using? (How are DVs operationalized?)
    - Why were these selected?

5) Does your research also involve factors or independent variables (IVs) (or interventions) that are expected to affect these DVs/criteria?

6) What are the factors/interventions/IVs of interest? How are these expected to influence or change what you want to impact?

    - If opportune, ask about immediate versus future generalizability: Are they testing factors that matter now (e.g., day vs. night operations) or that may matter in the future (e.g., interactive with one vehicle vs. interacting with a swarm of thousands of vehicles).

7) How would you determine the effects of the intervention/IVs?

    - How are the interventions/IVs operationalized? Why?
    - What research designs do you most often use? (Pre/post? Experiment? Surveys? Observation and ratings? Compare against past performance/gold standard?)
    - What comparisons groups do you usually use to evaluate effects of the intervention/IVs?

Section B: What the researcher may need to consider

"Going back to some of the points that you mentioned earlier . . . ?"

8) Are there other DVs/criteria that may be related to what you are interested in and which you are not measuring or including? (e.g., device to see through walls? Does that interfere with other head-mounted displays? Add too much weight?) Should you assess these additional DVs as well?

    - What are the barriers/obstacles to including those? (time, money, personnel, resources, equipment, difficult to measure, scope creep, etc. ?)

9) Are there other IVs that could mediate or moderate the IV–DV relationship that you are interested in?

- What are the barriers/obstacles to including those? (time, money, personnel, resources, equipment, difficult to measure, scope creep, etc.?)

10) Are there other IV–DV relationships that you should examine too (e.g., When I test high/low workload, should I always measure stress in addition to performance?)

11) Are the effects long-lasting? What is the timeframe of these effects?

12) Are there other extra effects that can emerge later? How would you know?

Section C: Additional questions

13) What are the local and global barriers to achieving your program's goals?

14) In an ideal situation, what constructs would you want to measure in your research and why? What measures would you use?

15) Compound tasks are tasks designed to force participants to engage with a tradeoff. For example, a detection task may be setup such that the participant would trade speed for accuracy of detection, or vice-versa. Compound tasks are considered more representative of real-world tasks than basic, generic experimental tasks. What compound tasks would you like to use in your research?

16) Compound tasks can have specific parameters and particular task requirements/demands. This may cause them to have less construct validity.

Possible example: Performance on a route-planning task operationalizes the construct of navigational ability of a Soldier. If there was an added factor of robot's aid, then operator's performance on the task may not reflect the construct navigational ability of the operator in the same way. But does it then operationalize the construct of navigational ability of the human–robot team? Human–robot trust and transparency too? To what degree would construct validity be a constraint in your research? (What theory would we be testing in this example? Resource theory to see how Soldier uses aid under low- and high-taskload situations? Or Multiple Resources Theory if the aid were given in different modality?)

17) What other compound measures would you like to use and why those? (e.g., surveys that may tap more than one construct/dimension).

18) What models (models/theories/frameworks) have you used in your research? Where do models fit into your program of research?

19) As you know, there are models at different levels (e.g., ACT-R (active control of thought–rational) is modeling at a micro level, whereas the resource theory and the situational awareness model are probably closer to the macro level. Which level of models are most relevant to your research and why?

20) Research can also be thought of in terms of 3 levels of analysis: physical, symbolic, and knowledge:

(i) Physical level: deals with physiological processes and anatomical circuits.

(ii) Symbolic level: deals with how the system operates and its function/what it does, involves the actual behavior of the system.

(iii) Knowledge level: deals with concepts like the system's goals, actions and its rationale, may not describe actual behaviors of the system.

Which level(s) is your research mostly at, and why?

21) How would you determine how effective a model/theory is in your research?

22) How much of your research is about trying to test model/theory?

23) How much of your research is about trying to develop model/theory?

List of Symbols, Abbreviations, and Acronyms

3-D	3-dimensional
4-D	4-dimensional
AAR	after-action review
APF	Assessment Planning Framework
APFT	Army Physical Fitness Test
app	application
ARL	US Army Research Laboratory
CAM	Conceptual Assessment Model
CBA	capabilities-based assessment
CSV	comma-separated value(s)
DIS	distributed interactive simulation
DOD	US Department of Defense
DV	dependent variable
HCE	Human Capabilities Enhancement
HMD	head-mounted display
HMT	human–machine teaming, or human–machine/–robot/–agent team
IR	infrared
IV	independent variable
LOS	line of sight
LTC	lieutenant colonel
MAA	Mobile Assessment Architecture
MOUT	military operations in urban terrain
NATO	North Atlantic Treaty Organization
NHST	Null Hypothesis Statistical Test
PoN	Point of Need

R&D	research and development
RDCT	Research Database Capture Tool
RVT	Research Visualization Tool
SME	subject matter expert
SNA	social network analysis
T&E	test and evaluation
ToM	theory of mind
TRADOC	US Army Training and Readiness Command
TTPs	tactics, techniques, and procedures
UI	user interface
UML	Unified Modeling Language
UMMPIREE	Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

2 DIR ARL
(PDF) IMAL HRA
RECORDS MGMT
RDRL DCL
TECH LIB

1 GOVT PRINTG OFC
(PDF) A MALHOTRA

1 ARL
(PDF) RDRL HRA A
C BURFORD

INTENTIONALLY LEFT BLANK.